

Tomasello, Mario Vincenzo; Burkholz, Rebekka; Schweitzer, Frank

Working Paper

Modeling the formation of R&D alliances: An agentbased model with empirical validation

Economics Discussion Papers, No. 2017-107

Provided in Cooperation with:

Kiel Institute for the World Economy – Leibniz Center for Research on Global Economic Challenges

Suggested Citation: Tomasello, Mario Vincenzo; Burkholz, Rebekka; Schweitzer, Frank (2017) : Modeling the formation of R&D alliances: An agentbased model with empirical validation, Economics Discussion Papers, No. 2017-107, Kiel Institute for the World Economy (IfW), Kiel

This Version is available at:

<https://hdl.handle.net/10419/172009>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



<http://creativecommons.org/licenses/by/4.0/>

Modeling the formation of R&D alliances: an agent-based model with empirical validation

Mario V. Tomasello, Rebekka Burkholz, and Frank Schweitzer

Abstract

The authors develop an agent-based model to reproduce the size distribution of R&D alliances of firms. Agents are uniformly selected to initiate an alliance and to invite collaboration partners. These decide about acceptance based on an individual threshold that is compared with the utility expected from joining the current alliance. The benefit of alliances results from the fitness of the agents involved. Fitness is obtained from an empirical distribution of agent's activities. The cost of an alliance reflects its coordination effort. Two free parameters a_c and a_1 scale the costs and the individual threshold. If initiators receive R rejections of invitations, the alliance formation stops and another initiator is selected. The three free parameters (a_c ; a_1 ; R) are calibrated against a large scale data set of about 15,000 firms engaging in about 15,000 R&D alliances over 26 years. For the validation of the model the authors compare the empirical size distribution with the theoretical one, using confidence bands, to find a very good agreement. As an asset of our agent-based model, they provide an analytical solution that allows to reduce the simulation effort considerably. The analytical solution applies to general forms of the utility of alliances. Hence, the model can be extended to other cases of alliance formation. While no information about the initiators of an alliance is available, the results indicate that mostly firms with high fitness are able to attract newcomers and to establish larger alliances.

(Published in Special Issue [Agent-based modelling and complexity economics](#))

JEL L14

Keywords R&D network; alliance; collaboration; agent

Authors

Mario V. Tomasello, Chair of Systems Design, ETH Zurich, Department of Management, Technology and Economics, Zurich, Switzerland

Rebekka Burkholz, Chair of Systems Design, ETH Zurich, Department of Management, Technology and Economics, Zurich, Switzerland

Frank Schweitzer, ✉ Chair of Systems Design, ETH Zurich, Department of Management, Technology and Economics, Zurich, Switzerland, fschweitzer@ethz.ch

Citation Mario V. Tomasello, Rebekka Burkholz, and Frank Schweitzer (2017). Modeling the formation of R&D alliances: an agent-based model with empirical validation. *Economics Discussion Papers*, No 2017-107, Kiel Institute for the World Economy. <http://www.economics-ejournal.org/economics/discussionpapers/2017-107>

1 Introduction

Collaboration can be widely observed in different social and economic systems, where agents strive to reach a common goal. Scientists collaborate to write joint publications (Katz and Martin, 1997), firms collaborate to file joint patents (Hoang and Rothaermel, 2005; Kim and Song, 2007), and software developers collaborate to create joint software products (Bitzer and Geishecker, 2010; Lakhani and Wolf, 2003). To *explain* collaboration, economic research has traditionally focused on different aspects of labor division (Durkheim, 2014) and productivity of teams (Scholtes *et al.*, 2016). However, in the wake of technology-driven economic growth, the question how to *boost* collaboration to foster knowledge transfer and innovation has become more important (Frenz and Ietto-Gillies, 2009).

The current research about the dynamics of R&D networks can be seen as a major contribution to better understand how firms collaborate in patenting activities. In this network representation, nodes depict the economic agents, i.e. the firms, and links between nodes their collaboration. Specifically, firms formally declare this collaboration in publicly announced *alliances*, which can involve more than two partners. So, it makes sense to ask how the *size of alliances*, i.e. the *number of partners involved*, can be explained by means of an agent based model, which is the aim of the current paper.

To address this questions, we can build on a number of empirical studies about R&D alliances. It was shown that, because firms are involved in different alliances at the same time, their collaboration results in a large network component, in which even firms not directly collaborating are still connected through other firms (See Figure 1). At the same time, a large number of small firm alliances exist that are not connected to the rest of the network. These co-existing sub-networks are called *components* in the following.

The formation of a strongly connected component can be seen as an emergent property of the economic network because it is not planned top down, but emerges during the process of alliance formation, if (some) firms become engaged in more than one alliance. Once such a strongly connected component exists, it greatly enhances the transfer of knowledge and the diffusion of innovations even between distant firms, so it is beneficial from a policy perspective.

(Tomasello *et al.*, 2014, 2017) have already proposed an agent-based model that is able to reproduce most of the properties of the observed R&D network. These properties include (i) the distribution of component sizes, i.e. the number of components of a given size plus the size of the largest connected component, (ii) the distribution of local clustering coefficients, i.e. the fraction of firms in a component that form triads (closed triangles) in their collaboration, (iii) the distribution of the lengths of shortest paths that connect any two firms in the network, and (iv) the distribution of degrees, i.e. the number of partners of a firm.

This agent-based model, while successfully reproducing network features along different dimensions, takes two empirical distributions as an input: (a) the distribution of agent's *activities*,

62 i.e. their propensity to engage into a collaboration, and (b) the distribution of *alliance sizes*,
63 i.e. the number of partners involved in an alliance. The latter has been investigated empirically
64 (see Hagedoorn, 2002; Tomasello *et al.*, 2016). Remarkably, one finds a broad and right-skewed
65 distribution of alliance sizes (see Figure 2). The same distribution was found even for different
66 industrial sectors (including manufacturing, research, financial and service sectors), that exhibit
67 substantial differences otherwise.

68 Previous modeling attempts in this field have, to the best of our knowledge, limited themselves
69 only to general features of the R&D network, such as the degree distribution or small world prop-
70 erties. With the current paper, we want to move the agent-based modeling one important step
71 forward, by *explaining* the distribution of alliance sizes as an emergent feature of an underlying
72 agent-based model instead of taking it from observations. This requires us to explicitly model
73 *how* agents form alliances, which implies to consider *why* agents form alliances. But given the
74 empirical work on alliance sizes, we have some *ground truth* to later judge the performance of
75 our agent-based model in reproducing the distribution of alliance sizes.

76 2 Empirical findings

77 2.1 The network of R&D alliances

78 **The dataset.** We build our empirical R&D network using the *SDC Platinum* database,¹ that
79 reports approximately 672,000 publicly announced alliances in all countries, from 1984 to 2009,
80 with a granularity of 1 day, between several kinds of economic actors (including manufacturing
81 firms, investors, banks and universities) for which we commonly use the term “firm” in the
82 following. We then select all the alliances characterized by the “R&D” flag; after applying this
83 filter, a total of $N = 14,829$ alliances, connecting $n = 14,561$ firms, are listed in the dataset.
84 An *R&D alliance* is defined as an declared partnership between two or more firms. This can
85 range from formal joint ventures to more informal research agreements, specifically aimed at
86 research and development purposes. Note that we do not have any information about the firm
87 that initiated the alliance, nor about the sequence in which firms joined an alliance.

88 The analysis of the data set, as well as all the network analyses and plots, are done by means of
89 the R software for statistical computing.²

90 **Reconstructing the collaboration network.** In the present study, we investigate the R&D
91 network aggregated over all years and all industrial sectors, which has to be reconstructed from
92 the data set. Firms are represented as nodes in the network and R&D alliances as undirected
93 links between nodes. Isolate nodes, i.e. firms not taking part in any R&D partnership, are simply
94 excluded from our network representation.

¹<http://thomsonreuters.com/sdc-platinum/>

²<http://www.r-project.org/>

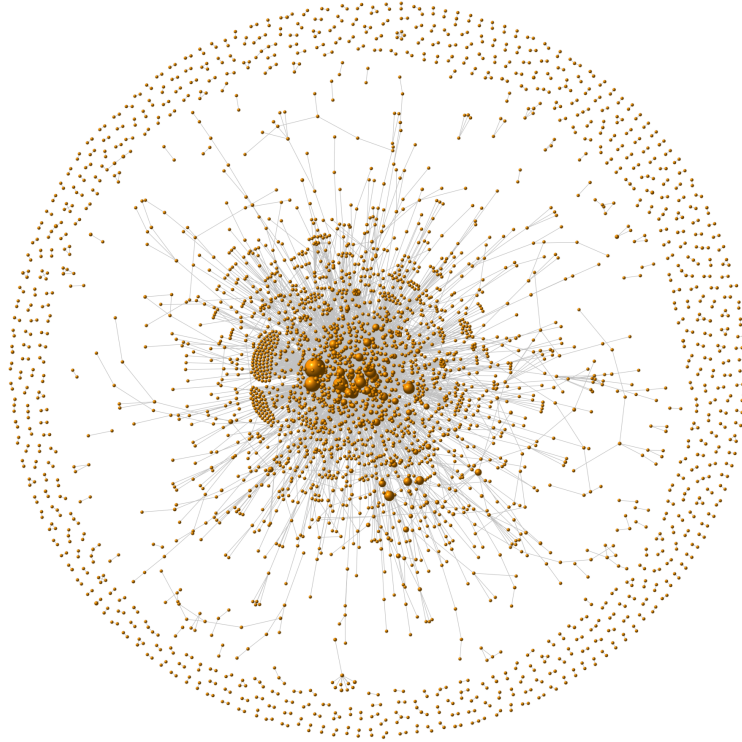


Figure 1: Visual representation of the R&D network that we analyze in this study – the size of the nodes encodes their fitness. We have used the *igraph* package (Csardi and Nepusz, 2006) and the Fruchterman-Reingold layout algorithm (Fruchterman and Reingold, 1991), which minimizes the number of crossing links.

95 When an R&D alliance involves more than two firms, we assume that all the corresponding
96 nodes are connected in pairs, forming a fully connected clique. A “standard” two-partner alliance
97 is then a fully connected clique of size 2. The choice of the fully connected clique – rather than
98 less interconnected network architectures – derives from the fact that alliances of more than two
99 partners, although representing only a minority, require great coordination and resources. There-
100 fore, they have to be associated with a higher number of links in the corresponding collaboration
101 networks. By following this procedure, the 14,829 R&D alliances listed in the dataset result in a
102 total of 21,572 links. The resulting network is shown in Figure 1.

103 **Distribution of alliance sizes** A salient feature of the R&D alliances in the SDC dataset is
104 the variable number of partners they involve. Most of the collaborations (93%) are stipulated
105 between two partners, the remaining ones involve three or more partners. In the following, we
106 denote by s the size of the alliance, whereas n indicates the number of firms and N the number
107 of alliances.

108 We report the empirical distribution of the alliance size, $p_s^e(s)$ in the R&D network in Figure 2.
109 As clearly visible, it spans one order of magnitude and is right-skewed. It should be noted that an

110 identification of the functional form of the distribution (e.g., power-law, exponential, log-normal
 111 and so on) is outside of the scope of this study. Our aim instead is to develop an agent-based
 112 model to reproduce this distribution, as described below.

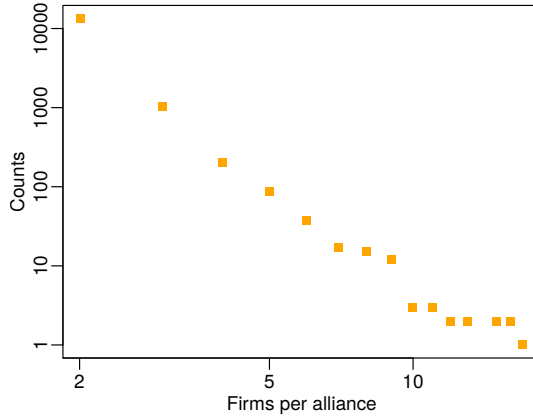


Figure 2: Histogram of the empirical alliance size distribution $p_s^e(s)$ measured on the R&D network. This distribution is later used to evaluate the outcome of our agent-based model.

113 2.2 Defining a fitness measure for agents

114 Our agent-based model requires an attribute, called “fitness”, that is assigned to each agent. It
 115 describes how attractive an agent is for the other agents, to form an alliance. To keep the model
 116 as a general as possible, we decide to proxy fitness by a measure which is *not* system specific, such
 117 as the operational value of a firm. We choose the so-called agents’ *activity* (Perra *et al.*, 2012),
 118 which has been already successfully used on various data sets, such as online microblogging, actor
 119 networks, R&D and co-authorship networks (Tomasello *et al.*, 2014). The empirical *activity* $\eta_{i,t}^{\Delta t}$
 120 of an agent i at time t , over a time window Δt , is defined as the *number of alliances* $n_{i,t}^{\Delta t}$ that
 121 involve agent i in the time window Δt ending at time t , divided by the total number of alliances
 122 $N_t^{\Delta t}$ involving *any* agent in the network during the same time period:

$$\eta_{i,t}^{\Delta t} = \frac{n_{i,t}^{\Delta t}}{N_t^{\Delta t}}. \quad (1)$$

123 It was found that activity distributions in most collaboration networks are right skewed and
 124 dispersed over several orders of magnitude, as in many other social and technological systems
 125 (Barabasi, 2005; Pastor-Satorras *et al.*, 2001). This is confirmed also for the case of R&D net-
 126 works, where the empirical activity values range from low 0.002 to the maximum value of 1.
 127 Applying this to the fitness of agents, this means that the agent with the highest fitness has a
 128 value of 2-3 orders of magnitude larger than the agents with the lowest fitness. Indeed, the vast

129 majority of the agents has a fitness equal to the minimum value, which is also the *median* value,
130 and the average fitness is only slightly higher than that. Only one agent has a fitness equal to 1
131 (the highest possible value).

132 Contrary to most network indicators that display strong variability and dependence on time,
133 especially in R&D networks (see Tomasello *et al.*, 2016), activity is a stable attribute that can
134 be assigned to firms to effectively estimate their propensity to engage in new collaborations, as
135 well as their attractiveness to potential new collaborators. Empirical activities are robust with
136 respect to (a) the time t at which they are measured, (b) the length of the selected time window
137 Δt , (c) the sectoral classification of firms or authors, as shown by Tomasello *et al.* (2014). Such
138 a stability makes activity a perfect empirical proxy for our fitness attribute.

139 Given the robustness with respect to the time window, we decide to compute the fitness values
140 using the longest possible window, i.e. the entire observation period, therefore $\eta_i \equiv \eta_{i,t=2009}^{\Delta t=26\text{years}}$.
141 This considers the full information from the data set and results in activities η_i that are always
142 strictly greater than 0 because, by definition, all firms in our network must be involved in at
143 least 1 alliance. In Figure 3 we report the *empirical distribution* of activity, i.e. of fitness, $p_{\eta}^e(\eta)$,
144 for the analyzed R&D network. Further, in Figure 1, we have used the empirical fitness values
145 of agents to scale their *size* in the collaboration network. Agents with higher fitness obviously
146 form the core of the empirical R&D network.

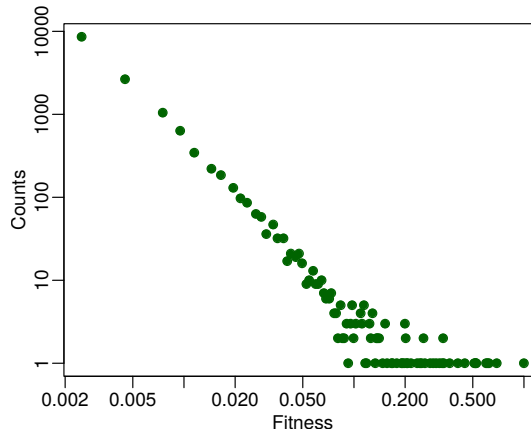


Figure 3: Histogram of agents' empirical fitness distribution, $p_{\eta}^e(\eta)$, measured on the R&D network. This distribution is later used as an input for our agent-based computer simulations.

3 The modeling approach

3.1 Agent-based model of alliance formation

In the following, we develop an agent-based model to reproduce the observed *size distribution* of consortia, shown in Figure 2. This distribution is the result of a dynamic process in which agents decide to initiate or to join an alliance, i.e. it can only be understood by modeling the growth of the collaboration network.

Fitness of agents and initiation of an alliance. Our model is a considerable extension of the network fitness model first proposed by Bianconi and Barabási (2001). Each agent i is assigned a *fitness* η_i which is fixed and independent of time. The values for the fitness are obtained from the empirical distribution $p_\eta^e(\eta)$, shown in Figure 3.

In our model, all n agents can become active with a *uniform probability*, which is chosen to be $1/n$, independent of their fitness. The sampling occurs with replacement, i.e. agents can also be chosen more than once to become active. Activity means here that an agent *initiates* a new alliance; hence we refer to her as the “initiator”. We do not have empirical information about the agent that initiated an alliance, hence the assumption of a uniform probability for the activation is reasonable.

For our simulations, we choose a discrete time t which measures the *time to form an alliance*. I.e. each time a new initiator is selected we start with $t = 0$ and the maximum time for alliance formation is denoted as T . The newly created alliance can grow only if new collaborators join. This process is reflected in two steps: (i) the initiator *invites* new collaboration partners, one per time step, (ii) the invitees *accept* or *reject* to join the alliance.

Utility of consortia. A number of agents form an alliance $\mathcal{C}(s^t)$ of size s^t which can change over time as new agents join the alliance. There can be many consortia of different sizes coexisting over time. The utility function, u^t , of the alliance combines the benefits, b^t , and the costs, c^t , of the collaboration of the s^t agents, i.e. both $b^t(s^t)$ and $c^t(s^t)$ depend on the current alliance size, s^t . Further, the benefits should be a monotonous function of the fitness values of the currently involved agents, i.e. $b^t(\dots, \eta_i, \eta_j, \dots)$, whereas the costs should reflect the coordination effort of the alliance and thus should be a monotonous function of the *size* of the alliance. For simplicity, we assume linear dependencies for the monotonous functions, i.e. the utility of an alliance is defined as

$$u^t = b^t - c^t; \quad b^t = \sum_{m=1}^{s^t} \eta_m; \quad c^t = a_c \cdot [s^t - 1] \quad (2)$$

where the parameter a_c allows to scale costs against benefits. We note that the costs scale with the number of alliance *partners* rather than with the number of their possible *connections*, which

170 would be quadratic, i.e. $s^t(s^t - 1)/2$, if an alliance is seen as a fully connected clique. The latter
 171 would account for a *superlinear* increase in the *coordination* effort between partners which sets
 172 strong limitations to larger consortia. Here, instead we assume some sort of *administration* cost
 173 based on the number of parties.

174 **Invitation of alliance partners.** As the alliance grows, its utility will change in a *non-*
 175 *monotonous* manner. Precisely, according to Equation 2, the utility will grow *only* if the fitness
 176 η_j of the new alliance member j is larger than the scaling constant a_c . To ensure that this
 177 condition is met, an initiator preferably invites agents with a high fitness. Precisely, similar
 178 to the fitness model of Bianconi and Barabási (2001), the initiator i chooses *potential* alliance
 179 partners j , one at a time, with a probability proportional to their fitness η_j .

180 Different from the mentioned model, it is however left to the agents to decide whether they want
 181 to accept this invitation, i.e. to join the alliance. The initiator repeats the invitation procedure
 182 until a number R of invited partners *refuse* the invitation to the alliance. I.e. R is a parameter
 183 of our agent-based model. If the *current number of rejections*, r^t , reaches R , we assume that the
 184 alliance is fully formed and stops to grow in size. At the same time the initiator loses its “active”
 185 status. If the initiator receives R rejections already from the first R selected partners, then no
 186 alliance is formed.

187 **Formation of collaboration links** The second step in the formation of the alliance is the
 188 decision of the invitee to accept, or to not accept, the invitation. An agent j decides to join an
 189 alliance \mathcal{C} at time $t + 1$ if the utility of the alliance, u^t is larger than a certain threshold, u_j^{thr} ,
 190 which is assumed to be *heterogeneous* across agents.

191 Specifically, we argue that the threshold of an agent to join an alliance *increases* with her fitness.
 192 The rationale behind this is that agents with a high fitness are very attractive for initiators of
 193 consortia and thus receive invitations very often. On the other hand, because of scarce resources,
 194 agents cannot simply accept all invitations, they have to be selective. Therefore, the higher the
 195 own fitness and the attractiveness for consortia, the higher the threshold to accept an invitation.
 196 Conversely, agents with a low fitness are not invited very often for an alliance, therefore they will
 197 be more inclined to accept invitations, i.e. their threshold is lower because of the lower fitness.
 198 Hence, it is reasonable to argue that $u_j^{\text{thr}} = a_l \eta_j$, i.e. u_j^{thr} is simply proportional to the fitness,
 199 where a_l is a parameter of the model, to be determined later.

This results in the following condition for agent j to join the alliance \mathcal{C} in the next time step

$$j \in \mathcal{C}_{t+1} \quad \text{if} \quad u^t \geq u_j^{\text{thr}} \quad \Rightarrow \quad \sum_{m=1}^{s^t} \eta_m - a_c[s^t - 1] \geq a_l \cdot \eta_j \quad (3)$$

200 **Implications.** Our agent-based model builds on an interesting tension between the *attractive-*
 201 *ness* and the *willingness* to become an alliance member, which is a novel point in the discussion

202 of fitness models. Hence, in our model the link formation process does *not* follow a simple prefer-
203 ential attachment rule. First, because agents do not *increase* their individual fitness by accepting
204 an invitation. Secondly, because agents become the more selective, the fitter they are.

205 It should be noted that, even though low-fit agents are more common in the network, high-fit
206 nodes agents a much higher chance to be selected as potential partners and to establish new
207 consortia. It is also clear from the setup of the model that agents with a low fitness would not be
208 able to establish a larger alliance. They can likely not attract agents with high fitness, nor can
209 they overcome the costs inclined in the formation of an alliance. Hence, larger consortia depend
210 on the initiation by agents with a high fitness and their ability to attract other agents with high
211 fitness.

212 Eventually, our agent-based model combines probabilistic elements, such as the activation of
213 agents and the invitation of partners, with deterministic elements, such as the decision of agents
214 to join the alliance. This decision differs from a “best response” rule, because agents do not decide
215 based on complete (or global) information about all existing consortia. Instead, they base their
216 decision only on the (local) information of the current offer.

217 Once the formation of an alliance is finished, this alliance is added to the existing network as a
218 clique, i.e. a fully connected cluster of size s^T , which is in line with our procedure to reconstruct
219 the network from empirical observations (see Section 2). This again differs from the mentioned
220 fitness model in that the network grows with the sequential addition of cliques, and not of single
221 links. The addition of a single edge linking two nodes, as argued in Section 2, can be thought of
222 as the addition of a fully connected clique of size 2.

223 3.2 Analytic description of the alliance size distribution

224 We now proceed in two different directions. First, we run stochastic simulations by implementing
225 the agent-based rules described above. Hence, at each time step we choose an agent to initiate
226 an alliance. Dependent on her success or failure, we add new cliques to the collaboration network
227 and continue with choosing a new agent in the next time step. This procedure is followed to
228 obtain the results discussed in the subsequent sections.

229 However, we also formalize the model in a more analytic way to obtain an expression for the
230 distribution of consortia sizes, $p_s(s)$. We start from the fitness distribution $p_\eta(\eta)$, which we take
231 as given from data. In accordance with the empirical distribution $p_\eta^e(\eta)$, we consider η as *discrete*
232 because of the binning given by the observations. We assume that $p_\eta(\eta)$ does not change during
233 the formation of consortia. That means even if agents with a given η_i have accepted an invitation
234 and can thus not be invited again to join the *same* alliance, we assume that the distribution $p_\eta(\eta)$
235 of the *remaining agents* is as before. This assumption is justified if it is unlikely that an alliance
236 grows to a significant proportion of the whole system, as it is the case in our studies.

237 With this, we derive an analytic proxy for the consortia size distribution $p_s(s)$. It saves consid-
238 erable computational effort and allows better insights into the model evolution. The distribution

239 $p_s(s)$ gives us the *probability* to find an alliance of *final size* s . This formation happened during
240 the time steps $t = 0, \dots, T$. At $t = 0$, an initiator is picked at random and thus has a fitness η_i
241 with probability $p_\eta(\eta_i)$. The alliance size at that time is $s^t \equiv s^0 = 1$. In each following time step,
242 another agent, which has the fitness η with probability $p_\eta(\eta)$, is invited to join the alliance. She
243 either accepts so that the alliance size increases, $s^{t+1} = s^t + 1$. Or she rejects which leads to an
244 increase in the number of rejections, $r^{t+1} = r^t + 1$. Hence, for the next time step $t + 1$ in the
245 alliance formation process always $t + 1 = s^t + r^t$ holds.

246 To evaluate the probability for both cases, we have to keep track of the alliance utility. For this,
247 we introduce a time-dependent distribution $p(t, s^t, r^t, b^t)$. It represents the joint probability that
248 at time step t an alliance has reached the size s^t , while offers were rejected r^t times.

249 b^t is the benefit of the alliance, i.e. the sum of the fitness values η of the alliance partners according
250 to Equation (2). The initial conditions for the time-dependent distribution are $p(0, 1, 0, b) =$
251 $p_\eta^e(\eta_i)$ because the alliance consists only of the initiator, and $p(0, s, r, b) = 0$ otherwise.

252 The arguments of $p(t, s^t, r^t, b^t)$ can only have the following values: the size ranges from $s^t \in$
253 $\{1, \dots, n\}$ with n as the total number of agents, the number of rejections from $r^t \in \{0, \dots, R\}$
254 where R is the maximum number of rejections that are still tolerated, and for the time $t \in$
255 $\{0, \dots, T\}$ where $T = n + R$ is the maximum time in which the formation of a single alliance is
256 possible. The benefit is bound to $b^t = \sum_{m=1}^{s^t} \eta_m \in [0, \sum_{m=1}^n \eta_m]$. Furthermore, only combinations
257 satisfying the condition $t + 1 = s^t + r^t$ are reasonable, because, in each time step, either s^t or r^t
258 are incremented. Otherwise, the growth of the alliance stops. Hence, $p(t, s^t, r^t, b^t) = 0$ can be set
259 for unreasonable combinations of t , s^t , and r^t .

260 Following these considerations, we start from the initial conditions given above and iteratively
261 deduce $p(t + 1, s^{t+1}, r^{t+1}, b^{t+1})$ from $p(t, s^t, r^t, b^t)$ at the previous time step. For this, we have to
262 take three different constellations into account: (1) the alliance *stops* growing at time t because the
263 maximum number of rejections R was reached, (2) the alliance could *potentially grow*, however
264 the invited agent *rejects* the offer, which leads to $s^{t+1} = s^t$ and $r^{t+1} = r^t + 1$, and (3) the
265 alliance *actually grows* because the invited agent *accepts* the offer, which leads to $s^{t+1} = s^t + 1$
266 and $r^{t+1} = r^t$.

267 For each of these three constellations we have to express the probability of its occurrence and the
268 fact that the boundary conditions for the given arguments t , s , r and b are met. For the latter, we
269 use the *indicator function*, which can be either zero or one. $\mathbb{1}_{[x,y]}(z) = 1$ means that the value of
270 z is within the range of x and y and $\mathbb{1}_{[x,y]}(z) = 0$ otherwise, whereas $\mathbb{1}_{[x]}(z) = 1$ means that the
271 value of z is precisely the value of x and $\mathbb{1}_{[x]}(z) = 0$ otherwise. This is more convenient and more
272 compact than **if** and **else** to express different cases. We need two indicator functions because
273 we have constraints on two variables, t and r^t , which determine $s^t = t - r^t$. With this, we can
274 capture the three different constellations as follows:

$$\begin{aligned}
p(t+1, s^{t+1}, r^{t+1}, b^{t+1}) &= \mathbb{1}_{[s^{t+1}+r^{t+1}, T]}(t+1) \mathbb{1}_{[R]}(r^{t+1}) p(t, s^{t+1}, R, b^{t+1}) \\
&+ \mathbb{1}_{[t+1-s]}(r^{t+1}) \mathbb{1}_{[1, R]}(r^{t+1}) p(t, s^{t+1}, r^{t+1} - 1, b^{t+1}) \left[1 - F_{p_\eta} \left(\frac{b^{t+1} - a_c(s^{t+1} - 1)}{a_l} \right) \right] \\
&+ \mathbb{1}_{[0, R-1]}(r^{t+1}) \sum_{\eta=0}^{\eta^*} p_\eta(\eta) p(t, s^{t+1} - 1, r^{t+1}, b^{t+1} - \eta)
\end{aligned} \tag{4}$$

275 The product $\mathbb{1}_{[s^{t+1}+r^{t+1}, T]}(t+1) \mathbb{1}_{[R]}(r^{t+1})$ in the first line of Eq. (4) refers to the case that
276 too many rejections have happened already and the alliance stopped growing. Hence, $b^{t+1} = b^t$,
277 $s^{t+1} = s^t$ and $r^{t+1} = r^t = R$. With this, $t+1 = s^t + r^t$, but still within the maximum time
278 allowed for alliance growth, $T = n + R$. $p(t, s^{t+1}, R, b^{t+1})$, on the other hand, gives the probability
279 of such a constellation at time t .

280 The second line of Eq. (4) counts all cases in which an invited agent rejects to join the alliance
281 because its fitness is too high in comparison with the alliance, $a_l \eta > b^t - a_c(s^t - 1)$, see Eq.
282 (3). The probability that this happens is given by the complementary cumulative distribution
283 function, $1 - F_{p_\eta}(\eta^*)$, with

$$F_{p_\eta}(\eta^*) = \sum_{\eta \leq \eta^*} p_\eta(\eta); \quad \eta^* = [b^t - a_c(s^t - 1)] / a_l \tag{5}$$

284 Because the rejection happens at time t , the current rejection value is $r^t = r^{t+1} - 1$, while the
285 size $s^{t+1} = s^t$ and the alliance benefit $b^{t+1} = b^t$ stay constant. The indicator function $\mathbb{1}_{[1, R]}(r^{t+1})$
286 ensures that r^{t+1} is still in the possible range of 1 (if that was the first rejection) and R . Otherwise,
287 this would have been captured in constellation (1). The second indicator function $\mathbb{1}_{[t+1-s]}(r^{t+1})$
288 just reflects the boundary condition $t+1 = s^{t+1} + r^{t+1}$. Because the agent has rejected the offer,
289 we have $s^{t+1} = s^t$.

290 The last line of Eq. (4) eventually considers all cases in which an agent accepts to join the
291 alliance at time $t+1$. In this case, $s^{t+1} = s^t + 1$, $r^{t+1} = r^t$ and $b^{t+1} = b^t + \eta$. The probability
292 of this occurrence has to be multiplied by the probability $p_\eta(\eta)$ to find agents of fitness η . The
293 summation goes over all possible benefit values, η , for which agents accept to join the alliance,
294 which follow from Eq. (3). This condition defines the value η^* introduced above. I.e., agents join
295 the alliance if their fitness η is between $[0, \eta^*]$. We can express this condition with respect to $t+1$
296 instead of t , i.e. $\eta^* = \{b^{t+1} - a_c(s^{t+1} - 2)\} / (1 + a_l)$.

297 The indicator function $\mathbb{1}_{[0, R]}(r^{t+1})$ eventually makes sure that the acceptable number of rejections
298 R is not already exceeded, i.e. the agent can still join the alliance. Otherwise, it had been
299 considered in constellation (1). We note that our approach also applies to more general forms
300 of cost functions $c^t(s^t)$ than just $a_c(s^t - 1)$. Only the bound η^* would need to be adjusted
301 correspondingly.

302 The iterations of Equation (4) occur until $t = T = n + R$ is reached, which is the maximum time
 303 possible for forming an alliance. This ensures that all consortia formations are counted in. The
 304 final alliance size distribution is then simply the marginal distribution

$$p_s(s) = \sum_{b^T} p(t, s^T, R, b^T) \quad (6)$$

305 This distribution depends, for a given number of agents, on the three parameters a_c , a_l , R and
 306 further on the empirical fitness distribution $p_\eta^e(\eta)$, which is given. Hence, a_c , a_l , R have to be
 307 determined in the following.

308 We emphasize that with our analytical solution we follow a probabilistic approach. That means,
 309 we take *all possible* constellation into account. This is equivalent to running a large number of
 310 computer simulations and averaging the results, at the end.

311 4 Results

312 4.1 Calibration of the agent-based model

313 Our first task is to calibrate the agent-based model introduced above. We take as model input
 314 the fitness distribution, $p_\eta(\eta)$, which is proxied by the empirical distribution $p_\eta^e(\eta)$ shown in
 315 Figure 3. It then remains to determine the set of parameters of the model, a_c to scale the costs
 316 of the consortium, Equation (2), a_l to scale the individual threshold for accepting invitations,
 317 Equation (3), and R which is the maximum number of rejections an initiator receives to stop
 318 the formation of an alliance. In order to determine these parameter values, we use a standard
 319 *maximum likelihood approach*. This can be based either on computer simulations of the agent-
 320 based model or on the numerical solution of the analytic expressions given in Equation (4). Both
 321 lead to the same results for all analyzed parameter combinations. Thus, we report the results
 322 from the numerics which can be obtained much faster.

323 We are interested in the distribution of the alliance size, $p_s(s)$, given the set of parameters
 324 (a_c, a_l, R) . In a first step, we have to re-normalize this distribution to alliance sizes $\tilde{s} \geq 2$ because
 325 we have no observations about $p_s^e(1)$. This renormalized distribution reads

$$p_{\tilde{s}}(\tilde{s}|a_c, a_l, R) = \frac{p_s(\tilde{s})}{\sum_{i=2}^n p_s(i)} \quad (7)$$

326 with $p_{\tilde{s}}(1|a_c, a_l, R) = 0$. The likelihood $\mathcal{L}(a_c, a_l, R)$ of each parameter combination (a_c, a_l, R) is
 327 then determined by the probability to observe our data which consists of N alliances with sizes
 328 s_1, \dots, s_N , given these parameters:

$$\mathcal{L}(a_c, a_l, R) = \prod_{i=1}^N p_{\bar{s}}(s_i | a_c, a_l, R) \quad (8)$$

329 For our simulations, as well as for the numerical solution of Equation (4), we will choose the set
330 of parameters $(\hat{a}_c, \hat{a}_l, \hat{R})$ that maximizes this likelihood:

$$(\hat{a}_c, \hat{a}_l, \hat{R}) = \arg \max \mathcal{L}(a_c, a_l, R) = \arg \max \mathcal{L}(a_c, a_l, R)^{1/N} \quad (9)$$

331 The exponent $1/N$ avoids the comparison of too small values and guaranties thus numerical
332 stability. The results are depicted in Figure 4. We note that there exists a region where all the
333 points with high goodness score are concentrated and that there is a sharp transition between
334 the red and the blue region. This region corresponds to definite values of $\hat{a}_c \simeq 0.04$ and $\hat{a}_l \simeq 2$,
335 but there is no definite value for R . In fact the “optimal” region is a line, along which R can vary
between 1 and 20. Thus, in the following we choose the maximum value $\hat{R} = 20$.

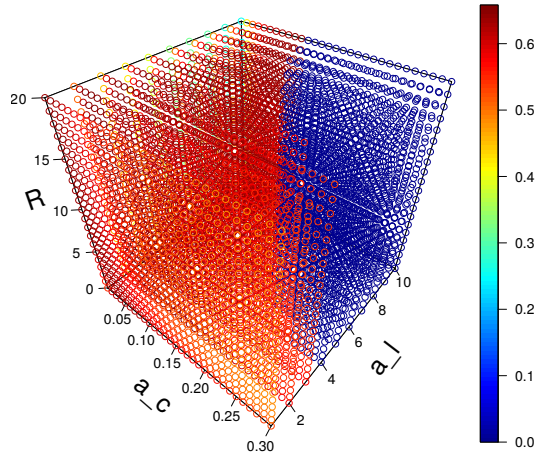


Figure 4: Results of the parameter values (a_s, a_l, R) using the maximum likelihood estimation, $\mathcal{L}(a_c, a_l, R)^{1/N}$.

336

337 4.2 Validation of the agent-based model

338 To challenge the validity of our agent-based model with the calibrated parameters, we need to
339 reject the Null hypothesis H_0 that the parameters $(\hat{a}_c, \hat{a}_l, \hat{R})$, are correct.

340 The assumption of our model is that all alliance formations are independent. This implies that
 341 our data has been generated by a multinomial distribution $M \sim Mult(N, p_{\tilde{s}}(2), p_{\tilde{s}}(3), \dots, p_{\tilde{s}}(n))$,
 342 where the probabilities $p_{\tilde{s}}(\tilde{s})$ result from our agent-based model. We use this multinomial dis-
 343 tribution to construct a region of 95% probability coverage, which we estimate by sampling 10^6
 344 times from M . In Figure 5 we represent this region as bands around the alliance size distribu-
 345 tion $p_s(s)$ obtained from the maximum likelihood approach. One band corresponds to the 0.025
 quantile, the second one to the 0.975 quantile for the probabilities $p_{\tilde{s}}(\tilde{s})$.

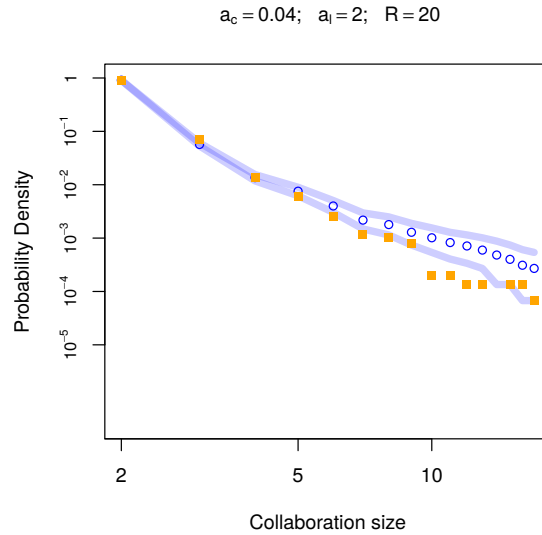


Figure 5: Alliance size distribution of the R&D network obtained from our agent-based model using the parameters from the maximum likelihood estimation. Orange squares represent the empirical size distribution, $p_s^e(s)$, blue circles the theoretical size distribution, $p_s(s)$ and blue lines the theoretical 95% quantiles obtained of the simulations of the multinomial distribution M . They define a confidence region for the theoretical size distribution.

346

347 A comparison with the empirical distribution $p_s^e(s)$ reveals the very good results. We see that in
 348 most of the empirical values are within the 95% confidence region. The four data points outside
 349 of that region are at least very close to the lower band. They also represent alliance sizes of only
 350 very small probabilities, about 10^{-4} , and thus cannot be considered as important deviations
 351 from the model. Hence, we conclude that our null hypothesis that the data are generated by the
 352 calibrated agent-based model *cannot* be rejected.

353 5 Conclusions

354 Our goal in this paper is to explain the empirically observed alliance size distribution, Figure 2,
 355 of R&D networks by means of an agent-based model that explicitly models the alliance formation
 356 process.

357 Our model builds on *heterogeneous* agents characterized by one individual parameter, their fitness
358 η_i . The distribution of fitness values is proxied by the empirical activity distribution, Figure 3,
359 as the only model input. Activity describes how often an agent was engaged in an alliance during
360 the observation period, which is 26 years in our case. This can be seen as an indication of the
361 agent’s attractiveness for other agents to collaborate with, and fitness should be interpreted in
362 the same manner. We have shown that the fitness distribution obtained this way is right skewed
363 and very broad.

364 Further, our agent-based model uses three free parameters that need to be determined in compar-
365 ison with empirical data. The calibration process is based on a maximum likelihood estimation
366 that returns those parameter values that match best the target, which is the empirical distribu-
367 tion of alliance sizes.

368 It is interesting to note that only two of these parameters, the scaling factors a_c for the cost of
369 the consortium and a_l for the individual threshold to accept an invitation obtain a stable value
370 in the maximum likelihood estimation, whereas the third parameter R , the number of rejections
371 to stop forming an alliance does not reach a definite value. Instead, we observe that equally
372 good likelihoods are obtained for a larger range of R between 1 and 20. Hence, our model works
373 *without* assuming a specific value of R . In other words, R can vary across time, industrial sectors
374 or even alliances without questioning the validity of our model.

375 For our model validation we used a high number of rejections, $\hat{R} = 20$. This is definitely realistic
376 for a system such as the global, inter-sectoral R&D network that we analyze. Here, firms have to
377 search for their partners among a huge number of potential candidates, making the establishment
378 of an R&D alliance potentially costly and risky. We argue that this leads to a very long and
379 cautious selection process, from the side of both the initiator and the invited firm. Therefore,
380 firms have to be willing to accept a high number of rejections, if they want to gain access to
381 external knowledge and eventually establish R&D collaborations with other firms.

382 Regarding the other two parameters, $\hat{a}_c = 0.04$ and $\hat{a}_l = 2$, we note from Equation (5) that
383 actually their ratio matters, as it determines the range of fitness values $[0, \eta^*]$ for which agents
384 join an alliance. The definite value of \hat{a}_c should be interpreted as rather large. I.e. when multiplied
385 with the size of the alliance, the cost in Equation (2) is rather high in comparison with the benefit
386 of the alliance, which is the sum of the fitness values of the agents. This has two consequences.
387 First, it restricts the maximum *size* of an alliance to values below 20. Second, it restricts the
388 maximum *number* of alliances with sizes larger than 2, because most agents in the system have a
389 rather low fitness and are thus not able to overcome the considerable cost of forming an alliance.
390 To illustrate this, an alliance of two agents with median fitness values exhibits a benefit of 0.004
391 and a cost of 0.04; or an alliance of four agents with median-fitness agents exhibits a benefit of
392 0.008 and a cost of 0.12, i.e. almost an order of magnitude larger.

393 This reflects the intention of our agent-based model. Agents with high fitness (typically incumbent
394 firms) are the ones that are most likely to receive an invitation. At the same time, they will most
395 often refuse the invitation, if an alliance consists of only agents with medium or low fitness nodes

396 (typically mid-size firms or startups). This leads to the high value of rejections obtained from
397 the maximum likelihood estimation. But if agents with a high fitness initiate an alliance, agents
398 with medium or low fitness are likely to accept this invitation. On the other hand, agents with
399 low fitness are not very selective to refuse any invitation because their threshold utility u_i^{thr} is
400 rather low, also as a consequence of the small value of \hat{a}_l .

401 The good match of our agent-based model with the empirical observations allows us to draw
402 some conclusions about the formation of real R&D alliances, for which no data is available.
403 As we have seen, alliances are more likely initiated by an incumbent firm of high fitness which
404 directs its interest toward a mid-size company or a startup. At the same time, the “bottleneck”
405 in establishing new alliances is probably on the initiator’s side, which has to take rejections and
406 keep looking for new partners until it finds the right one.

407 Our agent-based model was developed to reproduce the empirical distribution of alliance *sizes*.
408 One could be interested to know whether this model, using the parameter from the maximum
409 likelihood estimation, is also able to reproduce other features of the observed topology of the
410 R&D network. This is not the aim of the paper, but we can comment at least on the degree
411 distribution which was analyzed already by Tomasello *et al.* (2014). *Degree* refers to the *number*
412 of collaboration *partners* of an agent, not to the number of alliances the agent is involved. As such,
413 degree is not independent of the size of an alliance, and indeed the empirical degree distribution
414 was also shown to be right skewed and very broad.

415 However, we argue that the degree distribution *cannot* simply be obtained from our agent-based
416 model because this does not take *degree-degree correlations* into account. *Assortativity* reflects the
417 tendency of agents with *high* degree to form alliances with other agents with *high* degree, whereas
418 *dissortativity* would indicate that agents with *high* degree have the tendency to form alliances
419 with agents of *low* degree. Such degree-degree correlations have been detected by (Tomasello
420 *et al.*, 2016) both for sectoral R&D networks and for the aggregated R&D network used in this
421 paper. They play a role in particular for agents with high degree. Therefore, we can assume that
422 our agent-based model will be able to reproduce the right skewed and broad degree distribution,
423 but becomes increasingly worse in the range of larger degrees.

424 To conclude, our agent-based model provides a considerable step forward in identifying the real
425 mechanisms for alliance formation (Ahuja, 2000). In particular, with the distribution of alliance
426 sizes we are able to reproduce a feature that has received some attention in the existing literature,
427 but never a conclusive explanation. Our model can be used for stochastic agent-based simulations,
428 it also provides an analytical solution that considerably reduces the computational effort. We
429 emphasize that it is rather rare to obtain an analytic description of an agent based model. Our
430 derivations also apply to cases with different cost functions and are thus quite general. This
431 should inspire further agent based modelling approaches.

432 Our agent-based model is fully calibrated and validated against real data from the global interfirm
433 R&D network. It shows the emergence of a broad, right-skewed distribution of alliance sizes,
434 taking into account a heterogeneous fitness distribution of agents. On the methodological side,

our study provides an approach to infer the correct parameter values for the agent-based model, to interpret them and check their consistency with reality. Like for any agent-based model approach, we cannot conclude that our model is the only one able to explain and reproduce the alliance size distribution. However, the very good match with reality is a clear sign of plausibility for the set of agent rules that we propose, thus providing us with new insights into the micro dynamics of alliance formation.

Acknowledgements

M.V.T. and F.S. acknowledge financial support from the Swiss National Science Foundation, through grant 100014_126865, “R&D Network Life Cycles”. M.V.T. and F.S. acknowledge financial support from the Seed Project SP-RC 01-15 “Performance and resilience of collaboration networks”, granted by the ETH Risk Center of ETH Zurich. The authors thank Ryan Murphy for comments on an early version of this paper.

References

- Ahuja, G. (2000). The duality of collaboration: Inducements and opportunities in the formation of interfirm linkages. *Strategic management journal* **21**(3), 317–343.
- Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature* **435**(7039), 207–211.
- Bianconi, G.; Barabási, A. (2001). Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)* **54**, 436.
- Bitzer, J.; Geishecker, I. (2010). Who contributes voluntarily to OSS? An investigation among German IT employees. *Research Policy* **39**(1), 165–172.
- Csardi, G.; Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**(5).
- Durkheim, E. (2014). *The division of labor in society*. Simon and Schuster.
- Frenz, M.; Ietto-Gillies, G. (2009). The impact on innovation performance of different sources of knowledge: Evidence from the UK Community Innovation Survey. *Research Policy* **38**(7), 1125–1135.
- Fruchterman, T.; Reingold, E. (1991). Graph Drawing by Force-directed Placement. *Software-Practice and Experience* **21**(11), 1129–1164.
- Hagedoorn, J. (2002). Inter-firm R&D partnerships: an overview of major trends and patterns since 1960. *Research policy* **31**(4), 477–492.

- 466 Hoang, H.; Rothaermel, F. T. (2005). The effect of general and partner-specific alliance experience
467 on joint R&D project performance. *Academy of Management Journal* **48(2)**, 332–345.
- 468 Katz, J.; Martin, B. R. (1997). What is research collaboration? *Research Policy* **26(1)**, 1–18.
- 469 Kim, C.; Song, J. (2007). Creating new technology through alliances: An empirical investigation
470 of joint patents. *Technovation* **27**, 461–470.
- 471 Lakhani, K.; Wolf, R. G. (2003). Why Hackers Do What They Do: Understanding Motivation
472 and Effort in Free/Open Source Software Projects. *SSRN Electronic Journal* .
- 473 Pastor-Satorras, R.; Vazquez, A.; Vespignani, A. (2001). Dynamical and Correlation Properties
474 of the Internet. *Physical Review Letters* **87**.
- 475 Perra, N.; Goncalves, B.; Pastor-Satorras, R.; Vespignani, A. (2012). Activity driven modeling
476 of time varying networks. *Scientific Reports* **2**, 469.
- 477 Scholtes, I.; Mavrodiev, P.; Schweitzer, F. (2016). From Aristotle to Ringelmann: A large-scale
478 analysis of team productivity and coordination in Open Source Software projects. *Empirical
479 Software Engineering* **21(2)**, 642–683.
- 480 Tomasello, M. V.; Napoletano, M.; Garas, A.; Schweitzer, F. (2016). The Rise and Fall of R&D
481 Networks. *ICC - Industrial and Corporate Change* **26(4)**, 617–646.
- 482 Tomasello, M. V.; Perra, N.; Tessone, C. J.; Karsai, M.; Schweitzer, F. (2014). The Role of
483 Endogenous and Exogenous Mechanisms in the Formation of R&D Networks. *Scientific Reports*
484 **4**, 5679.
- 485 Tomasello, M. V.; Vaccario, G.; Schweitzer, F. (2017). Data-driven modeling of collaboration
486 networks: A cross-domain analysis. *EPJ Data Science* **6**, 22.

Please note:

You are most sincerely encouraged to participate in the open assessment of this discussion paper. You can do so by either recommending the paper or by posting your comments.

Please go to:

<http://www.economics-ejournal.org/economics/discussionpapers/2017-107>

The Editor