DISS. ETH Nr. 22448

Collaboration networks: their formation and evolution

A thesis submitted to attain the degree of
DOCTOR OF SCIENCES of ETH ZURICH
(Dr. sc. ETH Zurich)

presented by
MARIO VINCENZO TOMASELLO

MASTER IN GREEN MANAGEMENT, ENERGY AND C. S. R.	Bocconi University	ITALY
MASTER OF SCIENCE IN PHYSICS	University of Catania	ITALY

born on February 12, 1986

citizen of ITALY

accepted on the recommendation of
Prof. Dr. Dr. Frank Schweitzer
Prof. Dr. Hans Gersbach

2015

Acknowledgements

And so, here I am, writing the page that everyone will read first, but that comes after many months of hard work and lost sleep.

My first and most sincere thanks goes to Frank, who, with his experience, his explanations, his precise – and direct – comments, has guided and motivated me through this complicated, but full of satisfactions, path of my Ph.D. I learned from him many valuable lessons, not only in science.

My gratitude goes also to the rest of the Chair of Systems Design, the best working environment I have ever experienced. A place that can be challenging and stimulating, but at the same time relaxing and fun. Thanks to Claudio, for listening to my (not-so-standard) Italian every day and giving me an invaluable scientific help. Thanks to David, Antonios and Ingo because they were always there when I needed them. Thanks to my coworkers and friends René, Corneel, Pavlin, Vahan, Marcelo, Nico, Emre, Michelangelo, Simon, Nicolas and Yan, for supporting me in so many ways during the writing of my thesis, and also for sharing with me some memorable moments – both inside and outside the office. A special thanks goes to the girls of the Chair, Rahel, Rebekka and Adiya, who not only supported me, but also made my coffee breaks more pleasant day after day.

Thanks to my external collaborators and friends, Mauro and Nicola, for giving me the opportunity to learn, travel and enlarge my social circle. Without them, half of this dissertation would not be here.

Thanks to Heidi, for her constant encouragement, her sweet company, her delicious dinners, and for giving me the strength to go through many obstacles and difficulties. Thanks to Balint, for our bourgeois moments, and for teaching me some of the deepest and classiest quotes from English or American movies.

Thanks to all my friends in Zurich that have supported me in writing the thesis. Thanks to all the people that I have met during my trips around the world, for enriching my life in so many ways. Special thanks to Georgios, Michele, Gabriele and Riccardo; every moment on vacation and in the office with you is worth remembering.

A special *grazie*, from the bottom of my heart, goes to my parents, Orazio and Milena, and my little sister, Grazia. They have accompanied me all the days of my life, they rejoiced with me through the good times and supported me in the darkest hours, and the love I feel for them will never be affected by the miles and the mountains separating us. Without their dedication and their sacrifices, I would never be where I am now.

Thanks to my hometown friends, Alessio, Antonio and Michele. Together we grew up, we hit the ground running in many occasions, we laughed and cried at our own mistakes. But we always stood up and walked again, becoming the "magnificent four", united by a deep friendship that nothing can ever change.

Last but not least, a special thanks goes to the @IBM, for making me the person I am. I will be forever grateful to Alekampo, Benji and Ivano for all the moments we spent together, jumping from joy to despair, from play to study, from lucidity to madness, always trying to be critical and politically *in*correct. Moments that made me grow and that will keep us indissolubly united. An extraordinary thanks goes to Benji; our friendship has proven to be stronger than 10 years together as roommates, colleagues, trip companions and partners in crime.

Contents

Abstract

This thesis investigates the formation, evolution and performance of *collaboration networks*, with an emphasis on those networks in which every link-formation event involves a knowledge flow. Our work can be conceptually divided in three phases. First, we perform an extended *empirical analysis* of two prominent examples of such systems – namely R&D and co-authorship networks – to determine the microscopic rules for link formation and knowledge exchange between individual agents. Such rules include both network-endogenous and network-exogenous mechanisms. Second, we develop a set of *agent-based models* incorporating the interaction rules previously identified, that are able to reproduce a number of relevant observed features. We then validate such models against the empirical data. In doing so, we point out similarities and differences across sectors and research fields in R&D and co-authorship networks, obtaining – for the first time – a precise estimate of the relative weights of network-endogenous and -exogenous mechanisms. Third, by means of another agent-based model, we study how the microscopic rules for link formation affect the *performance* generated by these systems. We then validate this model against empirical R&D alliance and patent data, and investigate the optimality of such a real system with respect to its performance.

Remarkably, our framework is able to reproduce a large number of measures characterizing the network topology, including the distributions of degree, local clustering, path length and component size, as well as the emergence of network clusters. Furthermore, we find that endogenous mechanisms for link formation are predominant over the exogenous ones in most of the collaboration networks we study, thus supporting and quantifying the importance of existing network structures for selecting collaboration partners. With respect to the knowledge exchange phenomenon in a real R&D network, our models suggest that effective policies to obtain an optimized collaboration network would incentivize shorter R&D alliances and higher knowledge exchange rates than observed in reality.

Our results have an impact spanning from complex systems design to management science. Indeed, not only do we provide a unique methodology to systematically study link formation and knowledge exchange in dynamically evolving collaboration networks, but we also offer a procedure that allows to assess the performance of real systems and gives an indication on how to optimize them.

Kurzfassung auf Deutsch

Diese Dissertation untersucht die Entstehung und Entwicklung von Kollaborationsnetzwerken, sowie deren Performance. Ein besonderer Fokus wird auf jene Kollaborationsnetzwerke gelegt, in welchen die Linkerzeugung mit einem Wissensaustausch einhergeht. Zur Ermittlung der mikroskopischen Regeln, welche die Entstehung und Auflösung von Links zwischen Agenten bestimmen, führen wir zunächst eine empirische Analyse zweier prominenter Beispiele von Kollaborationsnetzwerken durch, nämlich R&D Kollaborationen und Koautorennetzwerke. Die auf diese Weise ermittelten Regeln beinhalten sowohl netzwerk-endogene wie auch -exogene Mechanismen. Als nächstes entwickeln wir agentenbasierte Modelle, welche die zuvor identifizieren Regeln nutzen, um eine Reihe relevanter, beobachteter Eigenschaften zu reproduzieren. Diese Modelle validieren wir anhand empirischer Daten und zeigen so Gemeinsamkeiten und Unterschiede zwischen verschiedenen Wirtschaftssektoren sowie Forschungsbereichen in R&D und Koautorennetzwerken auf. Hierdurch erhalten wir erstmals eine präzise Quantifizierung der relativen Gewichtung netzwerk-endogener und -exogener Mechanismen. Schliesslich nutzen wir ein weiteres agentenbasiertes Modell um zu untersuchen, wie die mikroskopischen Regeln zur Erzeugung von Links zwischen Agenten die kollektive Performance dieser Systeme beeinflussen. Wir validieren dieses Modell mittels empirischer Daten zu R&D Kollaborationen und Patenten und untersuchen wie optimal solch ein System in Bezug auf seine Performance ist.

Bemerkenswerterweise ist unser Framework in der Lage eine grosse Zahl charakteristischer Netzwerkmasse wie bspw. Knotengradverteilung, lokaler Clusterkoeffizient, Pfadlängen, oder die Grösse verbundener Komponenten, sowie die Entstehung von Clustern zur reproduzieren. Darüber hinaus zeigen unsere Ergebnisse, dass in den meisten der von uns untersuchten Kollaborationsnetzwerke endogene gegenüber exogenen Mechanismen dominieren, ein Resultat welches nicht nur die Bedeutung von Netzwerkstrukturen bei der Auswahl von Kollaborationspartnern hervorhebt sondern sie auch quantifiziert. Hinsichtlich einer Optimierung des Wissensaustauschs in echten R&D Netzwerken, legen unsere Modelle Anreizstrukturen nahe welche, verglichen mit den tatsächlich beobachteten Systemen, kürzere Allianzen mit einer höheren Rate des Wissensaustauschs fördern.

Unser Ansatz stellt eine neue Methode zur systematischen Untersuchung von Linkerzeugung und Wissensaustausch in dynamischen Kollaborationsnetzwerken dar. Da wir diese Methode mit einem Verfahren koppeln, welches es erlaubt, die Effizienz eines Systems zu bewerten und zu optimieren, reicht die Bedeutung unserer Ergebnisse vom Design komplexer Systeme bis hin zu den Managementwissenschaften.

Sintesi in italiano

Questa tesi tratta della formazione, evoluzione e prestazioni di “reti di collaborazione”, con particolare riferimento a quelle reti in cui ogni creazione di un legame collaborativo comporta un flusso di conoscenza. Il nostro lavoro può essere concettualmente diviso in tre fasi. In primo luogo, eseguiamo un’analisi empirica approfondita di due importanti esempi di tali sistemi – reti di coautoraggio e reti di Ricerca e Sviluppo (R&S) – per determinare le regole microscopiche che portano alla formazione di collaborazioni tra singoli agenti. Tali regole comprendono meccanismi sia endogeni che esogeni rispetto alla rete stessa. In secondo luogo, sviluppiamo una serie di “modelli ad agenti” che incorporano le regole di interazione precedentemente identificate, e che sono in grado di riprodurre una serie di caratteristiche empiriche. Tali modelli sono poi convalidati con dati reali, in modo da evidenziare analogie e differenze tra settori e/o campi di ricerca, nelle reti di R&S e di coautoraggio, ottenendo – per la prima volta – una stima precisa dei pesi relativi dei meccanismi di rete endogeni ed esogeni. In terzo luogo, per mezzo di un altro modello ad agenti, studiamo come le regole microscopiche per la formazione delle collaborazioni influiscono sulla prestazione generata da questi sistemi. Questo modello sarà convalidato con dati su alleanze di R&S e brevetti, e servirà ad analizzare le prestazioni di questo sistema reale.

Sorprendentemente, i nostri modelli sono in grado di riprodurre un gran numero di misure che caratterizzano la topologia delle reti analizzate, comprese le distribuzioni di grado, di raggruppamento locale, di percorsi di rete e di dimensioni delle componenti connesse, nonché la comparsa di comunità nella rete. Inoltre, troviamo che i meccanismi di rete endogeni sono predominanti su quelli esogeni nella maggior parte delle reti di collaborazione analizzate, a sostegno dell’importanza delle strutture di rete esistenti per la selezione di nuovi partner di collaborazione. Per quanto riguarda il fenomeno dello scambio di conoscenza in una rete reale di R&S, i nostri modelli suggeriscono che delle politiche mirate ad ottenere una rete ottimizzata dovrebbero incentivare alleanze di R&S più brevi e velocità di scambio di conoscenza più elevate di quanto osservato nella realtà.

I nostri risultati hanno un impatto che va dalla progettazione di sistemi complessi fino al management. Infatti, non solo forniamo una metodologia unica per studiare sistematicamente la formazione di collaborazioni e lo scambio di conoscenza in reti che evolvono dinamicamente, ma offriamo anche una procedura che permette di valutare le prestazioni di sistemi reali e dà al tempo stesso un’indicazione su come ottimizzarli.

Summary

For Chapters 2 to 6, a detailed summary is presented on the first page.

Chapter 1: Introduction introduces the thesis, by explaining its rationales, briefly reviewing the existing empirical and theoretical literature, and describing the used data and methodology. We also list the precise research questions addressed by this dissertation.

Chapter 2: Stylized facts of R&D networks reports an exhaustive empirical analysis of R&D networks and tracks their evolution in a large number of industrial sectors – including manufacturing, services and public research sectors – over a long time period (1986-2009). We evaluate the temporal and sectoral robustness of many statistical properties of real R&D networks, by examining a complete set of indicators, to the best of our knowledge larger than any previous empirical study. We also investigate the formation of R&D alliances from a microscopic point of view, by means of an econometric model. We use a novel approach, in which the observation unit is not the firm, but every potential pair of firms, and the dependent variable is the formation of an alliance.

Chapter 3: Similarities among collaboration networks extends our study of network trends and patterns on the domain of co-authorship networks in scientific disciplines. We identify all the differences, but also the similarities, across collaboration networks in the R&D and the co-authorship domains. We find that some features are indeed universal and robust. In particular, the size of collaboration events, the agents' *activity* (i.e. their propensity to engage in a collaboration), and the presence of structural communities in the network (that go beyond the agents' sectoral or geographical positions).

Chapter 4: Modeling the formation of collaboration networks incorporates the building blocks and microscopic rules identified in Chapters 2 and 3 into an agent-based model including both network-endogenous and network-exogenous mechanisms for link formation. Remarkably, by fitting only some macroscopic network properties, our model is able to reproduce a number of *microscopic* measures characterizing the network topology, including the distributions of degree, local clustering, path length and component size, and the emergence of network clusters. By validating the model on both R&D and co-authorship networks, we find that network-endogenous mechanisms are predominant over the exogenous ones in most of the collaboration networks we investigate. Therefore, we precisely quantify the importance of existing network structures for selecting new collaboration partners in different domains.

Chapter 5: Modeling the exchange of knowledge in a collaboration network investigates the phenomenon of knowledge exchange in a dynamic collaboration network, by means of a second agent-based model. The model allows us to study the complex interdependencies and mutual fedbacks between the network structure and the nodes'

intrinsic characteristics (i.e. their knowledge basis). We define the *performance* of the collaboration network as the distance travelled by all of its agents in a metric knowledge space. The model parameters we investigate are the link rewing rate of the network and the agents' interaction radius. We find that, depending on the parameter values, the agents tend to cluster around one or a few attractors in the knowledge space, whose position is an emergent property of the system. And – more importantly – we find that there exist optimal values for both the link rewiring rate and the agents' interaction radius to maximize the network performance.

Chapter 6: Towards a more general modeling framework combines the two agent-based models developed in Chapters 4 and 5 into a unified agent-based model, that we validate on empirical alliance and patent data for R&D networks. The underlying knowledge space we consider in our real example is defined by IPC patent classes, allowing for a precise quantification of every firm's knowledge position. Such unified framework is able to predict the topology of the emerging collaboration network and the effect that this has on the firms' patenting activities, as well as providing indications for an improved R&D alliance network. Precisely, we find that the real R&D network does not maximize the distance travelled by its agents in the underlying knowledge space. Effective policies to obtain an optimized collaboration network – as suggested by our model – would incentivize shorter R&D alliances and higher knowledge exchange rates than observed in reality.

Chapter 7: Discussion and conclusions lists the main findings of this thesis and the impact they have in complex systems design and management science. We argue that the main contributions of the thesis are: (i) a unique methodology to systematically study link formation and knowledge exchange in dynamically evolving collaboration networks, and (ii) a procedure that allows to assess the performance of real systems and gives an indication on how to optimize them. Finally, we outline future research directions in the field of optimal system design and systemic risk in collaboration networks.

List of publications

The present dissertation is based on the following publications:

- Mario Vincenzo Tomasello, Moritz Müller, and Frank Schweitzer. "Innovator Networks". In: *Encyclopedia of Social Network Analysis and Mining.* Springer, New York, pp. 737–742 (2014).

- Mario Vincenzo Tomasello, Nicola Perra, Claudio Juan Tessone, Márton Karsai, and Frank Schweitzer. "The Role of Endogenous and Exogenous Mechanisms in the Formation of R&D Networks". *Scientific Reports*, 4, 5679 (2014).

- Mario Vincenzo Tomasello, Mauro Napoletano, Antonios Garas, and Frank Schweitzer. "The Rise and Fall of R&D Networks". arXiv:1304.3623 (2013). Submitted to *Industrial and Corporate Change*, current state: under revision.

and the following unpublished working papers:

- Mario Vincenzo Tomasello, Claudio Juan Tessone, and Frank Schweitzer. "Network dynamics and the exploration of knowledge in collaboration networks".

- Mario Vincenzo Tomasello, Claudio Juan Tessone, and Frank Schweitzer. "The effect of R&D collaborations on firms' technological positions".

Additional results, not included in the dissertation, are available in:

- Antonios Garas, Mario Vincenzo Tomasello, and Frank Schweitzer. "Selection rules in alliance formation: Strategic decisions or abundance of choice?" *arXiv:1403.3298* (2014).

Chapter 1

Introduction

> "Any existing structures and all the conditions of doing business are always in a process of change. Every situation is being upset before it has had time to work itself out. Economic progress, in a capitalist society, means turmoil."
>
> Joseph Alois Schumpeter
> *Capitalism, Socialism and Democracy* (1942)

The famous economist and political scientist Joseph Alois Schumpeter devoted only six pages to the process of "creative destruction" in his book *Capitalism, Socialism and Democracy*, in which he described capitalism as the "perennial gale of creative destruction". However, it has been argued (Cox and Alm, 2008) that this concept has become the basis for modern thinking on how economies evolve.

Indeed, we do believe that the creation of something new, be it a successful innovation, a brilliant invention, or just a piece of knowledge, naturally involves a huge amount of learning and uncertainty. Among the other words, Schumpeter (1942) uses the term "new combinations" to indicate the driving forces leading to economic growth: new products, new methods of production, new sources of supply, exploitation of new markets, new ways to organize business. However, independently of its specific purpose, we argue that every process of "new combination" exhibits an intrinsically dynamic and collaborative nature.

Such recombinant processes have been observed in many real systems, spanning from social interactions (Garcia *et al.*, 2015, 2014a) to online communities (Garcia *et al.*, 2013), from research and development activities (Ahuja, 2000b; Hagedoorn, 2002) to scientific production (Börner *et al.*, 2003; Sarigöl *et al.*, 2014), from finance (Battiston *et al.*, 2013) to global trade (Schweitzer *et al.*, 2009). The collaborative nature of these processes causes the emergence of networks in the totality of the above mentioned systems, often exhibiting a complex and time-evolving structure.

The outcome of these systems is not simply given by the sum of all agents' outcomes, but depends on the set of their interactions, which – in extreme cases – can generate unintended and unpredictable effects (think, for instance, of cascading failures in interbank networks, Fig. 1.1). The structure of these systems is almost never engineered; on the contrary, it emerges spontaneously from the consecutive interactions of a great number of agents. This is why the understanding of the microscopic rules for link formation (and destruction) is an unavoidable step to effectively intervene on their macroscopic outcome.

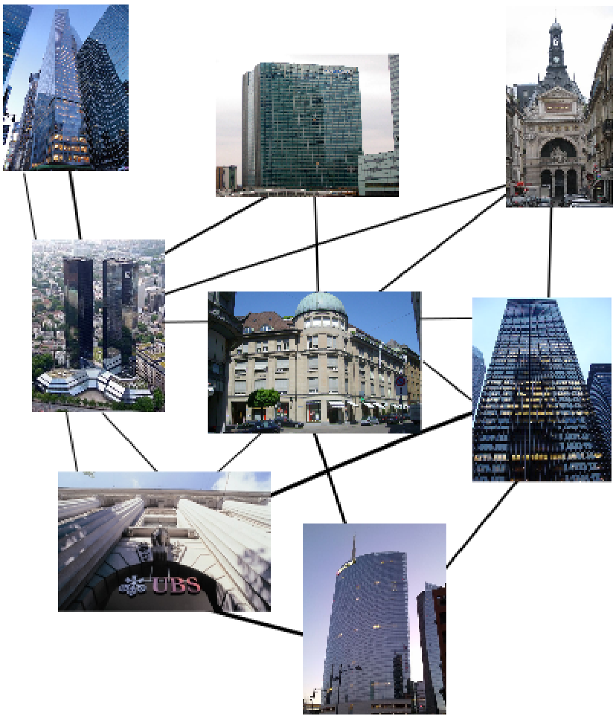


Figure 1.1: A free representation of an interbank network. Source: systemic risk as emerging phenomenon (Burkholz, 2014).

1.1 The emergence of collaboration networks

Some examples of networked systems that have received a great deal of attention in academic research and the media are social networks, real and virtual, human and non-human (Garcia *et al.*, 2015, 2014b), infrastructural networks, i.e. power grids, transportation, the Internet (Albert *et al.*, 2004; Gonzalez *et al.*, 2008; Pastor-Satorras and Vespignani, 2007) and financial or global trade systems (Caldarelli *et al.*, 2013; Schweitzer *et al.*, 2009), because their impact on our daily life is clearly visible and relatively easy to study in a quantitative fashion.

However, we argue that there exist other networked systems that significantly contribute to economic growth, whose formation and evolution is not fully understood yet, probably because both their short-term and long-term effects are difficult to quantify. We refer to *collaboration networks* emerging in economic systems and in scientific research, aimed at the creation of new *knowledge*. The present dissertation wants to fill this gap, by studying the structure of such collaboration networks, the microscopic rules leading to link formation and dissolution between individual agents, and how they affect the aggregated outcome generated by these systems. In addition, we want to investigate whether it is possible to optimize real systems with respect to such outcome.

When we use the terms "collaboration networks", we refer to those networks in which every event of link formation involves a certain flow of knowledge. Precisely, we will be focusing in this dissertation on how existing network structures affect the establishment of new collaborations and the joint production of new pieces of knowledge – such as scientific papers or patents – and not on the subsequent steps of invention (technical feasibility) or innovation (economic success).

1.1.1 Performance and systemic risk

Financial or social systems are not the only ones exhibiting a high degree of connectedness among agents. As we show later in the present dissertation, collaboration networks are also characterized by significantly high connectedness, often showing the emergence of a unique, giant network component (see Fig. 1.2 for a real example). This means not only that the outcomes of all agents may exhibit complex interdependencies, but also that a possible failure affecting one agent may be propagated and amplified through the network – for instance, in the case reported in Fig. 1.2, various counterparty risks might spread to several industrial sectors.

Moreover, the majority of such systems exhibit a dynamical structure, being their links continuously formed, terminated or rewired – in Fig. 1.3, we report an example of dynamic collaboration network, where a strongly interconnected *core* co-evolves together with a sparse and volatile *periphery*. In order words, collaboration networks show at the same time a certain degree of robustness *and* adaptation to change. A deep investigation of both the link formation mechanisms and the way they affect knowledge flows through the network is crucial to design efficient collaboration networks and minimize their vulnerability.

Recent works in economics (Battiston *et al.*, 2012; Kaushik and Battiston, 2013) have shown that being densely connected is not always beneficial for a system: certain networks have proven to be "too interconnected to be stable". In the same line, other works in temporal network theory (Pfitzner *et al.*, 2013; Scholtes *et al.*, 2014) have proven that the connectedness properties and the temporal order of the links heavily affect the network's performance, in terms of diffusion efficiency or other dynamical processes.

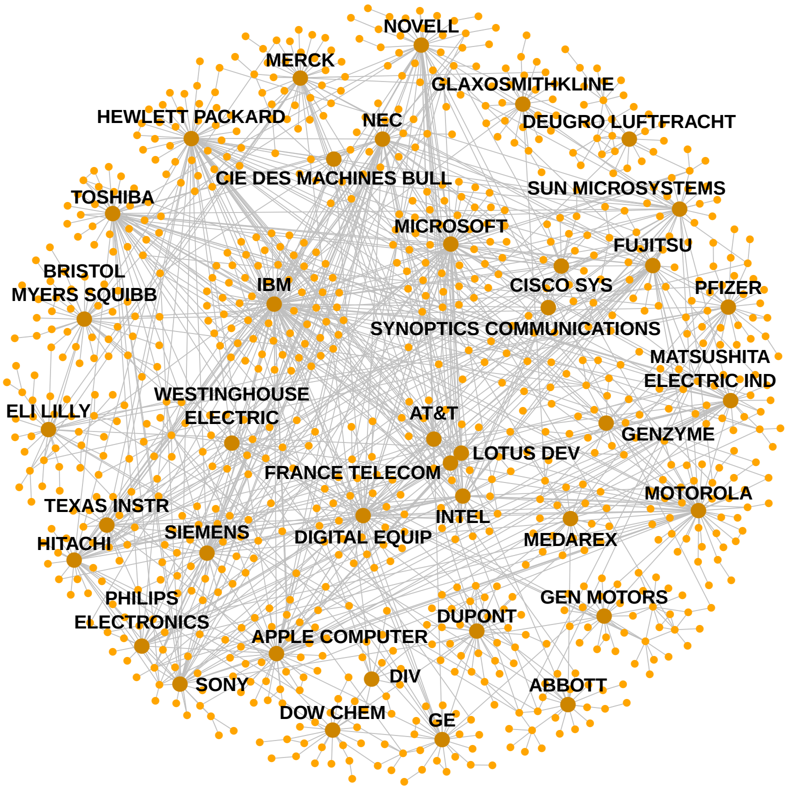


Figure 1.2: Visual representation of the global R&D network obtained from the Thomson Reuters SDC alliance dataset. Every node represents a firm and every link an R&D alliance. We depict the 30 core firms and their respective circles of influence.

Like every other system exhibiting high interconnectedness, we argue that also collaboration networks are exposed to risks deriving from cascading processes or consecutive performance drops spreading on the network (Vespignani Alessandro, 2010), that we indicate as "systemic risk". In particular, dynamical collaboration networks involving measurable knowledge flows through their links – such as inter-firm R&D networks or co-authorship networks – are especially susceptible to this problem.

Precisely, we believe that systemic risk can potentially materialize in two ways. The first way is that the network grows with a particular topology that does not result in an efficient knowledge spreading, because it is either too sparse or too dense – this is directly related to the presence of structural holes in the network (see Burt, 1992; Kleinberg *et al.*, 2008; Vega-Redondo and Goyal, 2007, for more examples). The second way is that – even though the network allows optimal knowledge flows in a given period of time – it is not resilient to shocks, which can result in a sudden drop of performance – or a total network extinction – if some of the external conditions change. A remarkable example is represented by the collapse of an entire online social network, that has been studied in Garcia *et al.* (2013), or the collapse of apps, websites or other software projects,[1] studied in Ribeiro (2014).

We want to stress that such a risk is not exogenous to the system, but it endogenously emerges as a consequence of the system structure and functioning. Differently from a risk represented by a single disruptive event to a company (such as an earthquake or a financial collapse), that can be relatively easily quantified and monetized, systemic risk is endogenously present in the system, and exhibits complex dependencies not only on the agents' features, but also their interactions.

Given that collaboration networks in most cases cannot be designed with a *top-down* approach, we highlight the importance to understand how a collaboration network forms and evolves from a microscopic point of view. This way, through a *bottom-up* approach, collaboration networks can be improved and steered towards configurations that are not only more efficient in spreading knowledge, but also more resilient to shocks.

1.1.2 Complex structure

In order to characterize our time-evolving collaboration networks, we use the tool of *complex networks*. We argue that, despite the variety of actors taking part in such processes, their treatment can universally be abstracted to the study of networks in which links between agents represent their collaborations. This set of relationships enables the agents to coordinate their efforts and create new knowledge or explore new knowledge trajectories. Like many other instances of socio-economic networks (Barabasi, 2005; Barabasi and Albert, 1999; Pastor-Satorras *et al.*, 2001; Powell *et al.*, 2005), we argue that all collaboration networks are characterized by three key factors:

Agents are heterogeneous. Collaborating agents are diverse, both with respect to their nature (organizational types span from multinational firms to single scientists) and to their knowledge (each actor is endowed with a unique knowledge base).

Diversity fosters interaction. Agents tend to collaborate because their knowledge or skills are complementary. Provided that some pre-conditions for the interaction exist, the diversity among players represents one of the most important incentives to collaboration.

Networks are self-organized. Knowledge-based networks are typically not designed but emerge spontaneously. Agents maximize their utility when forming and dissolving relationships. They are, in turn, affected in many ways by their position in the network. This creates a mutual feedback between the actors and the network structure, resulting in path-dependence and co-evolution.

We intend to investigate the formation of such networks, quantifying the effects of two different aspects: previous network structures (that we also refer to as *network-endogenous*

[1] Information about dead projects is available at `http://techcrunch.com/tag/deadpool/`

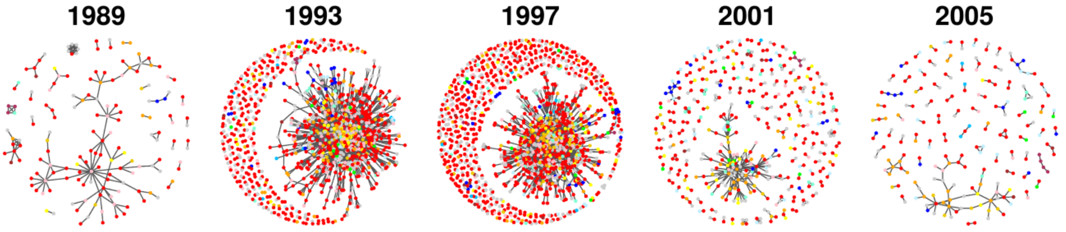


Figure 1.3: Evolution of the computer software R&D network, obtained from the Thomson Reuters SDC alliance dataset. The emergence of a giant connected component is clearly visible from this representation. After the mid-nineties, the collapse of such component is associated with the growth of a sparse and weakly connected peripheral component.

mechanisms) and other agent-specific factors (that we also refer to as *network-exogenous mechanisms*). Differently from previous studies, here we do not investigate the formation of spatial clusters of firms and scientists, or the optimization of intra-organizational knowledge production, or the knowledge transfer from one system to another.

This thesis will investigate the *formation* and the *dissolution* of collaborations in several domains, attempting to quantify the microscopic rules leading to the macroscopic features that we observe in a number of real cases, spanning from peculiar network topologies to the emergence of temporal patterns. Next, we will quantify the dependence of some network *performance indicators* on a set of microscopic network parameters, thus paving the way to the design of optimal collaboration networks.

1.1.3 Two prominent examples: R&D and co-authorship networks

Following the identification of the microscopic interaction rules, based on our systematic empirical observations, we will develop a set of agent-based models that are able to reproduce the *dynamics*, the *structure* and the *outcome* of the observed collaboration networks. The validation of the models, together with the fine tuning of the relevant parameters, will give us insights into the optimality of the analyzed empirical systems, while providing at the same time indications on how to improve them.

We will thoroughly analyze two prominent examples of real collaboration domains, namely inter-organizational R&D networks and co-authorship networks in science. This choice has been made mainly for two reasons: the impact that such systems have on human development (Gersbach *et al.*, 2013), and the availability of extensive data, that allow for a quantitative and rigorous analysis. Based on theoretical arguments such as Schumpeter's idea of innovation as a recombination process, or the resource-based view of the firm, companies can be considered as the fundamental units aimed at creating innovation in an economic system (Nonaka and Takeuchi, 1995). On the other hand, basic scientific research – as testified by its unprecedented growth in the last decades (Liu *et al.*, 2005) – drives most of the human development, fueling our technological progress (Salter and Martin, 2001).

R&D collaboration networks. The domain that we study first (and most extensively) in this dissertation is represented by inter-firm Research and Development (R&D) alliances. A considerable amount of literature has been developed specifically about collaborating firms. Besides, companies-related data sources, such as databases on strategic alliances and joint ventures, offer the possibility to construct large, often longitudinal networks, allowing extensive empirical studies. This is exactly the focus of Chapter 2 of the present dissertation.

The 1980s and 1990s witnessed an unprecedented growth of R&D alliances (Hagedoorn, 2002). This has been investigated by two different streams of empirical literature.[2] One body of contributions studies the salient features of empirically observed collaboration networks (see e.g. Ahuja, 2000a; Fleming *et al.*, 2007; Hanaki *et al.*, 2010; Powell *et al.*, 1996, 2005; Roijakkers and Hagedoorn, 2006). These studies have mainly found that collaboration networks tend to be small worlds characterized by short path lengths and high clustering (Watts and Strogatz, 1998). In addition, they tend to be highly heterogeneous and centralized, although there exist some differences across industries (Rosenkopf and Schilling, 2007).

A second body of work studies the relation between network features and firm performance (Cowan and Jonard, 2004; Letterie *et al.*, 2008), both at company and aggregate level. One still open debate is whether dense interconnections are more conducive to knowledge diffusion than weak bridging ties between separate communities (Granovetter, 1985, 1983). Indeed, clusters of densely connected firms foster collaboration efforts by generating trust, punishment of opportunistic behaviors, and common practices (Ahuja, 2000b; Coleman, 1988; Walker *et al.*, 1997). Conversely, by creating a structural hole in the network, firms have access to different sources of knowledge spillovers, economizing on the costs of direct collaborations (Burt, 1992; Rowley *et al.*, 2000). Other works (Gulati, 1995b; Gulati and Gargiulo, 1999; Rosenkopf and Padula, 2008) pointed out the relation between a firm's position in the network and its knowledge base. It has been found that two players should not be too similar nor too different in their knowledge bases in order to engage in a collaboration (Cohen and Levinthal, 1990, 1989; Lazer and Friedman, 2007).

In particular, we now briefly describe two illustrative examples from the empirical litera-

[2] See Cohen (1995), Powell and Grodal (2006), Walker (2005), Ebers (1997) and Veugelers (October 1998) for a more extensive overview.

ture. In the first one (Rosenkopf and Schilling, 2007), the comparison of alliance networks across industries highlights how technology relates to network structures. The alliance network for 32 industrial sectors has been analyzed in terms of size, connectivity, centralization, small-world properties and other network-related measures. As we will also show in the continuation of the present thesis, the networks exhibit different structures across industries, mainly determined by their respective technological features. Technological dynamism and separability of innovation are positively related to the share of firms participating in alliances (thus influencing the size of alliance network) and to the average number of alliances formed by each firm (thus influencing the average degree). Moreover, concentration of architectural control is positively related to the asymmetry in the number of alliances (thus influencing the dispersion of the degree distribution) and to the appearance of small world properties in the alliance network (high clustering and short path lengths).

The second work (Powell *et al.*, 2005) studies the evolution over time of the alliance network in the commercial field of the life science industry. Using panel data on biopharmaceutical alliances, the factors that drive alliance formation have been investigated. It has been found that the network structure is determined by both past alliance activity and intrinsic characteristics of the agents. Four mechanisms of link creation, used by different agents in different periods, have been identified: (a) *accumulative advantage*: the most connected agents receive a disproportionate share of new links; (b) *homophily*: new partners are chosen on the basis of their similarity to previous partners; (c) *follow-the-trend*: agents show a herd-like behavior; (d) *multi-connectivity*: agents choose partners that connect to one another through multiple independent paths.

This dissertation will *combine* the two approaches described in the above illustrative examples. We will, as a first step, identify the microscopic rules leading to alliance formation, by means of an econometric model and a set of complex network tools. Then, we will incorporate them into an agent-based model to reproduce the emergence of the macroscopic network structures observed in real R&D networks.

Co-authorship networks. The second domain of collaboration networks that we examine is represented by co-authorship networks in science, i.e. networks of scientific authors whose links constitute co-authored papers. Price (1965) was one of the first scholars suggesting to use the scientific method to study science itself. Since then, research in bibliometrics and scientometrics has developed tools to analyze more and more extensive publication datasets. A great number of works focus on identifying networks or clusters of authors, papers, or references, providing “maps” of science (e.g. Boyack *et al.*, 2005; Leydesdorff, 1987; McCain, 1991). One prominent example of this stream of literature is Newman (2004b), who has compared the co-authorship networks in three different science fields: biomedical research, physics and mathematics. All the fields proved to be similar, in terms of broad degree distributions and assortativity coefficient, but different with respect to their mean degree or clustering coefficient.

Alternative methods based on co-word analysis were later developed to identify semantic themes (Callon *et al.*, 1983). Recent progress in complex networks, as well as visualization techniques, have recently lead to advanced representations of knowledge domains (Börner *et al.*, 2003). Moreover, advances in computing capabilities have facilitated the analysis of large-scale datasets; for instance, Bollen *et al.* (2009) used clickstream data to provide a high-resolution and up-to-date view of scientific activity, correcting the underrepresentation of social sciences and humanities that is commonly found in citation data.

Furthermore, a second literature stream deals with models reproducing the growth of collaboration network in science (e.g. Banks and Carley, 1996; Snijders, 2001). In particular, Börner *et al.* (2004) introduce a model called TARL (for topics, aging, and recursive linking) that grows at the same time co-authorship and paper citation networks. The model incorporates a partitioning of authors and papers into topics, a bias for authors to cite recent papers, and a tendency for authors to cite papers cited by papers that they have read. Given its scope, the present thesis will contribute especially to the latter stream of literature, by investigating the formation mechanism of collaborations in science, rather than to the characterization and mapping of research fields.

1.2 Contribution of the present study

The dissertation is structured in seven Chapters. The current Chapter introduces the thesis, by explaining its rationales, briefly reporting the empirical and theoretical background, and describing the used data and methodology. An extended empirical study follows: Chapter 2 reports a thorough analysis of R&D networks, in several industrial sectors, and includes an econometric model investigating the microscopic rules for alliance formation. In Chapter 3, we extend the analysis to a set of co-authorship networks in scientific disciplines, and identify all the similarities across domains, which will constitute the building blocks for an agent-based model, thus concluding the empirical part of the thesis.

Chapter 4 develops an agent-based model that is able to reproduce the topology of different observed collaboration networks; the model is validated against real data, on both R&D and co-authorship networks. In Chapter 5, we explore the mechanisms of knowledge exchange in a dynamic network by means of a second agent-based model, identifying a theoretical optimal prescription to maximize the aggregate agents' knowledge exploration. Chapter 6 combines the two agent-based models into a unified agent-based model, that we validate on empirical alliance and patent data for R&D networks; such unified framework is able to predict the topology of the emerging collaboration network and the effect that this has on the firms' patenting activities, as well as providing indications for an efficient, improved R&D alliance network. Finally, Chapter 7 concludes and gives an overview of the open questions left by the present study and the future research directions.

1.2.1 Research Questions

As we have already mentioned, the present thesis is divided in two main parts, an empirical one and a modeling one. None of the two parts, or none of the thesis chapters, alone, are able to answer specific research questions, but only the combined empirical and modeling effort allows us to address the following relevant questions:

- **RQ1.** What are the individual rules for link formation in R&D networks? Are we able to validate them empirically by means of an econometric model?

- **RQ2.** To what extent is the formation of R&D collaborations driven by the agents' position in the R&D network? And to what extent is it driven by other, network-unrelated factors?

- **RQ3.** Can we extend to co-authorship networks in science the questions about R&D networks, provided that we consider authors as the collaborating agents and co-authored papers as links of the network?

- **RQ4.** We expect the presence of a different incentive scheme. How does this change the individual rules and, consequently, the structure and the dynamics of the corresponding collaboration network?

- **RQ5.** Are we able to develop an agent-based model using the interaction rules derived through our empirical study, that is able to reproduce the dynamics and the structure of the observed collaboration networks?

- **RQ6.** Building upon this agent-based model and tuning one or more of its parameters, are we able to find an optimal collaboration dynamics – i.e. maximizing some aggregate indicator of knowledge production?

The extensive data collection and processing – carried out in Chapters 2 and 3 – will allow us to quantify the effect of technological position and social embeddedness of companies and scientific authors on the formation of collaborations, thus addressing RQ1, RQ2 and RQ3. The individuation of the microscopic interaction rules (in Chapter 3) and their implementation into an agent-based model (in Chapter 4) allow us to answer RQ2, RQ4 and RQ5.

The development of the knowledge exchange agent-based model (Chapter 5) complements the answer to RQ5, giving at the same time a preliminary answer to RQ6. Finally, the validation and the fine tuning of the parameters in our general modeling framework (Chapter 6) allows us to fully address RQ6.

For the sake of completeness, it has to be mentioned that we have investigated one additional research question, namely the possibility to identify common behaviors – in terms of network centrality evolution – for the most successful agents, i.e. those having the highest knowledge production. We have decided not to include the subsequent results in this thesis, because this question, being centered on the behavior of single agents, lies outside the broader scope of the dissertation, which is the investigation of the formation and evolution of collaboration networks as a whole. However, our investigation has generated a paper including an empirical analysis and a simple agent-based model that is able to explain the agents' centrality evolution in a real R&D network. The results are available in Garas *et al.* (2014).

1.2.2 Data

This thesis will make use of four different datasets, conveniently disambiguated and merged. The first dataset is the *SDC Platinum* alliance database, provided by Thomson Reuters.[3] This dataset reports all publicly announced R&D partnerships, from 1984 to 2009, between several kinds of economic actors (including manufacturing companies, investors, banks and universities). All the data have been handled and processed through a PostgreSQL data server. A total of 14,829 alliances are listed in the SDC dataset, with their beginning date and a short description of the alliance purpose. Every company is associated with its official name, a short business description and a SIC (Standard Industrial Classification) code, allowing us to assign each firm to the right industrial sector.

The second data source that we use to quantify companies' knowledge production is the Patent Citations Data by NBER (the U.S.A. National Bureau of Economic Research).[4] The dataset contains detailed information on about 3 million patents granted in the U.S.A. between 1974 and 2000. Every patent is associated with one or more assignees and with an IPC (International Patent Classification) class. Companies are associated with a unique identifier, and a relatively big part of them are also matched to the Compustat dataset, containing financial information about all firms traded in the U.S. stock market. A significant amount of work will be devoted to merge the NBER patent dataset with the SDC alliance dataset.

The third dataset is a list of all papers and citations within the American Physical Society

[3] `http://thomsonreuters.com/sdc-platinum/`

[4] `http://www.nber.org/patents/`

(APS) domain,[5] including all papers' title, authors, affiliation, research field, and so on. The covered journals are Physical Review Letters, the Reviews of Modern Physics, and all the Physical Review journals, for the period 1983-2010. The fourth and last dataset is the Microsoft Academic Search (MSAS) dataset,[6] that we employ to obtain disambiguated information about all authors' first and last name, author's address and e-mail address, institution, department, city. More details about all datasets will be given in the continuation of the thesis.

All our data are gathered and organized in PostgreSQL databases.[7] The analysis of the databases, as well as regression models and plots, are done by means of the R software for statistical computing.[8] Agent-based models are developed in the C and Python programming languages. The visualization of the networks is done through the *i-graph* package for R (Csardi and Nepusz, 2006).

1.2.3 Methods

The general approach that we adopt throughout the thesis is *data driven modeling.* This means that an extended empirical analysis constitutes the starting point of our study. This analysis allows us to identify a set of regularities and similarities across collaboration networks, that will be used as building blocks for the subsequent theoretical models.

Next, we incorporate the identified blocks as microscopic rules of several agent-based models, that – through computer simulations – are aimed at reproducing the observed network topology and other features of real collaboration networks. We argue that the use of agent-based models is the most appropriate approach to perform this task, in that it allows to abstract the constituents of many systems (and their properties) into self-sufficient agents and to impose rules of interaction among them (Schweitzer, 2007, Chap. 1).

Moreover, the use of agent-based models reflects the conceptual approach of complex systems: it is only through the interaction of many individual elements that the emergent properties of such systems can be understood. In particular, we will focus on two emergent properties: the resulting network topology and some indicator of aggregate knowledge production.

As a final step – and in line with our data driven modeling approach – we will validate our models against empirical data and fine tune the values of the most relevant model parameters. It should be noted that the goal we intend to achieve with our study is *not* an accurate prediction of the system outcomes (unlike weather forecasting or other

[5] `http://www.aps.org/`

[6] `http://academic.research.microsoft.com/`

[7] `http://www.postgresql.org/`

[8] `http://www.r-project.org/`

engineering applications). This means that we will not encode the real system into our agent-based models by including as much detail as *possible*. We will rather try to identify the minimum set of agent attributes and interaction rules that can reproduce a certain emergent behavior, including as much detail as *necessary*. Therefore, based on empirical findings, we start from the simplest feasible set of rules, and compare the emerging outcome of the model against available data. We then add complexity step-by-step, until the desired level of detail is reached or the selected macroscopic effect is successfully reproduced.

Finally, an added value of this approach is the possibility to compare the *optimality* of the real systems with the simulated ones, in terms of some appropriate performance indicators. A complete exploration of the parameter space can give us a useful indication on how real system can be improved in sub-optimal cases.

Chapter 2

Stylized facts of R&D networks

Summary

In this Chapter we carry out an exhaustive empirical analysis of R&D networks. Drawing on a large database of publicly announced R&D alliances, we track the evolution of R&D networks in a large number of industrial sectors – including manufacturing, services and public research sectors – over a 25-year time period (1986-2009). Our main goal is to evaluate the temporal and/or sectoral robustness of many statistical properties of real R&D networks. We examine a complete set of indicators, larger than any previous empirical study, to the best of our knowledge, thus providing a complete description of R&D networks. We find that most network properties are invariant across sectors. In addition, they do not change when varying the scale of aggregation (pooled or sectoral) at which the network is observed. This represents a first step towards the identification of universal patterns in collaboration networks. Moreover, for the specific case of R&D networks, we find that most indicators are characterized by a rise-and-fall dynamics, with a peak in the mid-nineties. Finally, we investigate the formation of R&D alliances from a microscopic point of view, by means of an econometric model. We use a novel approach, in which the observation unit is not the firm, but every potential pair of firms, and the dependent variable is the formation of an alliance. We find that previous network structures, along with potential network structure changes, determine the alliance formation as much as the network-unrelated variables, i.e. firm country, sector and technological knowledge basis. However, a model including both network-related and -unrelated variables has the highest possible goodness of fit in explaining the formation of R&D alliances.

Based on M. V. Tomasello, M. Napoletano, A. Garas, and F. Schweitzer, "The Rise and Fall of R&D Networks", arXiv:1304.3623 (2013). Submitted to *Industrial and Corporate Change*, current state: under revision. M.V.T. contributed to designing the research questions, produced all the statistical analyses and the plots, and wrote most of the manuscript. The section containing the econometric model, not included in the original manuscript, was specifically designed and written by M.V.T. for the present dissertation.

2.1 Theoretical and empirical background

The increasing importance of R&D collaborations for industrial innovation has originated both empirical and theoretical research on R&D networks. he empirical works have tried to shed light on the structural properties of R&D networks, by showing that R&D networks are typically sparse and characterized by heavily asymmetric degree distributions (e.g. Hagedoorn, 2002; Hanaki *et al.*, 2010; Powell *et al.*, 2005; Rosenkopf and Schilling, 2007). Furthermore, R&D networks display "small world" properties (e.g. Fleming *et al.*, 2007; Fleming and Marx, 2006).

The theoretical studies have shown that R&D collaborations allow innovation either via resource sharing (Goyal and Joshi, 2003; Goyal and Moraga-Gonzalez, 2001; Westbrock, 2010) or via the recombination of firm's knowledge stock with those of its partners (Cowan and Jonard, 2004; König *et al.*, 2011, 2012). One key prediction of these theoretical models is that – under non-negligible costs of collaboration – R&D networks should be organized as core-periphery architectures, i.e. they should display a core of densely connected firms, in turn linked with a periphery of firms having few alliances among them. Nevertheless, to the best of our knowledge, no empirical study has tried so far to confirm or deny the presence of core-periphery architectures in R&D networks, nor study the evolution of other network indicators on a variety of industrial sectors.

The analysis that we develop in the current Chapter contributes to the foregoing empirical and theoretical literature along several dimensions.

First, we analyze the R&D networks in a large number of manufacturing and service sectors. After analyzing the *pooled R&D network*, i.e. the network containing all alliances independently of the sectors to which the partners belong, we study a series of R&D networks for several industrial sectors at a 3-digit SIC level. Via this disaggregated analysis, we are able to check whether the network properties that have been analyzed by the current literature for sectors like computers (e.g. Hanaki *et al.*, 2010) or pharmaceuticals (e.g. Powell *et al.*, 2005) are robust across different sectors of activity. In addition, by comparing the properties at the pooled and at the sectoral levels, we are able to check for the presence of *universal* properties of R&D networks that hold irrespectively of the scale of aggregation at which they are observed.

Second, we perform a longitudinal analysis of empirical R&D networks. In particular, we consider the network dynamics in the period from 1986 to 2009. This procedure allows us to check whether network properties are robust over time, or if instead they exhibit different trends in different time-periods.

Third, we investigate a broad set of network properties. We start our analysis by studying the basic network measures that have so far been considered in the empirical literature (size, degree heterogeneity, small world property).

Fourth, we study the formation of R&D collaborations through a novel econometric approach. In our study, the observation unit is not a single firm, but every potential pair of firms in the network. Our dependent variable is the formation of a link. The independent variables include i. firm structural characteristics (that is, country, industrial sector and technological knowledge basis), ii. current network centrality and other embeddedness measures, iii. potential change in network centrality if the considered link is actually formed.

2.2 Data and methodology

In this Section, we present the dataset upon which we base the empirical analyses of the current chapter. Together with the description of the data, we provide a detailed explanation of the methodology that we employ to build an R&D network and to compute the relevant network measures, coefficients and distributions. This methodology will be used in most of the following chapters of the present dissertation.

We define an *R&D network* as a representation of the research and development alliances occurring between firms in one or more industrial sectors in a given period of time. Such network consists of a set of *nodes* and *links* connecting pairs of nodes. In our representation, each node of the network is a *firm* and every link represents a *R&D alliance* between two firms. By R&D alliance, we refer to an event of partnership between two firms, that can span from formal joint ventures to more informal research agreements, specifically aimed at research and development purposes. To detect such events, we use the *SDC Platinum* database, provided by Thomson Reuters, that reports all publicly announced alliances, from 1984 to 2009, between several kinds of economic actors (including manufacturing firms, investors, banks and universities). We then select all the alliances characterized by the “R&D” flag; after applying this filter, a total of 14,829 alliances are listed in the dataset.

Information in the SDC dataset is gathered only from announcements in public sources, such as press releases or journal articles. Nevertheless, despite the bias that could be introduced by such a collection procedure, Schilling (2009) shows that the SDC Thomson dataset provides a consistent picture with respect to alternative databases (e.g. CORE and MERIT-CATI) in terms of alliance activity over time, geographical location of companies and industry composition.

Because the SDC Platinum dataset does not have a unique identifier for each firm, all the associations between alliances and firms (i.e. the construction of the network itself) are based only on the firm names reported in the dataset. Thus, it could happen that two or more entries are listed with different names, because they appear in two distinct alliances, even though they correspond to the same firm. For this reason, we check all firm names and control for all legal extensions (e.g. "ltd", "inc", etc.) and other recurrent keywords (e.g. "bio", "tech", "pharma", "lab", etc.) that could affect the matching between entries referring to the same firm. We decide to keep as separated entities the subsidiaries of the same firm located in different countries. The raw dataset contains a total of 16313 firms, which are reduced to 14561 after running such an extensive standardization procedure.

In our network representation, we draw a link connecting two nodes every time an alliance between the two corresponding firms is announced in the dataset. An alliance is associated with an *undirected* link, as we do not have any information about the initiator of the alliance. When an alliance involves more than two firms (*consortium*), all the involved firms are connected in pairs, resulting into a fully connected clique. Following this procedure, the 14,829 alliance events listed in the dataset result in a total of 21,572 links. Similarly to Rosenkopf and Schilling (2007), the R&D network we consider in our study is *unipartite*, as we only have one set of actors ("the firms"), whose elements may be connected – or not – by publicly announced alliances.[1]

Multiple links between the same nodes are in principle allowed (two firms can have more than one alliance on different projects). Nevertheless, as we aim at studying the connections between firms, and not the number of alliances a firm is involved in, we discard this information and use *unweighted* links in our network representation. For this reason, we define the *degree* of a node as the number of other nodes to which it is linked, i.e. the number of partners that a firm has – not the number of alliances. Furthermore, a firm appears in the R&D network only if it is involved in at least one alliance. Our study is focused exclusively on the embeddedness of firms into an alliance network. For this reason, isolated nodes are not part of our network representation.

Both the links and the nodes of the R&D network are characterized by an entry/exit dynamics. Alliances between firms have a finite duration (see Deeds and Hill, 1999; Phelps, 2003). This causes some firms to disappear from the network, after they no longer participate in any alliance. Likewise, many new firms that were not listed in any previous alliance may enter the network at the beginning of a new year. Our longitudinal study clearly requires precise temporal information about the formation and the deletion of alliances. The SDC Platinum dataset contains the beginning date of every alliance, but there is no information about any of the ending dates (firms do usually not organize press releases to announce the end of an alliance). We are thus forced to make some assumptions about the alliance durations. We start by drawing the duration of every alliance from a normal distribution with mean value from 1 to 5 years and standard deviation from 1 to 5 years, and we find that all our results remain qualitatively unchanged by changing the

[1] Our work differs from previous empirical studies (e.g. Cantner and Graf, 2006; Hanaki *et al.*, 2010; Lissoni *et al.*, 2013) which construct the network through the association of firms with patents and/or inventors. Those studies use patent data to build the network and associate elements in the set "firms" to the elements in the set "patents". This way, the network they obtain is *bipartite*.

mean value and the standard deviation within these ranges. More precisely, the variation of the standard deviation has nearly no influence on the patterns exhibited by of all measures we compute on the networks. The variation of the mean alliance duration changes the absolute values of the network indicators, but it does not affect their time-evolution and peak positions. Given the strong robustness of the R&D network to the variation of alliance lengths, we take a conservative approach and assume a fixed 3-year length for every partnership, consistently with previous empirical work (e.g. Deeds and Hill, 1999; Phelps, 2003; Rosenkopf and Schilling, 2007). More precisely, we link two nodes when an alliance between the corresponding firms occurs and we delete this link 3 years after its formation. In this way, we are able to build 26 snapshots of the R&D network – one for every year – from 1986 to 2009. From now on we call the network containing all companies, irrespective of their industrial sector, the *pooled R&D network*.

Every firm listed in the SDC Platinum dataset is associated with its SIC (Standard Industrial Classification), a US-government code system for classifying industrial sectors. This allows us to build the *sectoral R&D networks* for the several sectors that we identify in the dataset. A sectoral R&D network centered around a given sector contains only alliances in which at least *one* of the partners has a three-digit SIC code matching the selected sector (see also Rosenkopf and Schilling, 2007, for a similar approach). The rules for link deletion are the same as in the pooled R&D network. More precisely, we select for our study the 30 largest industrial sectors, in terms of number of firms engaged in alliances in 1995 (the year in which the pooled R&D network reaches its maximum size). This list includes manufacturing and service sectors. It has to be noticed that the latter includes also sectors like "laboratories and testing companies" and "universities". Table 2.1 provides the list of the different sectors we consider in our study.

2.3 Key network measures and trends across sectors

In the present Section, we provide a detailed empirical characterization of both the pooled R&D network and the sectoral R&D networks, by computing a set of network indicators along the whole observation period. The results of our analysis are grouped into five subsections: basic network statistics, heterogeneity in alliance behavior, assortativity, small world and communities, core-periphery structures.

2.3.1 Basic network statistics

We start our analysis by presenting a set of fundamental network indicators, such as size and density, as well as visual representations of R&D networks. Fig. 2.1 shows six snapshots of the pooled R&D network. The plots are produced using the *igraph* library[2] for *R*, and the networks are displayed using the Fruchterman-Reingold algorithm (cf. Fruchterman and Reingold, 1991). This is a force-based algorithm for network visualization which positions the nodes of a graph in a two-dimensional space so that all the edges are of similar length and there are as few crossing edges as possible. The result is that the most interconnected nodes are displayed close to each other in the two-dimensional plot. The ten largest industrial sectors are depicted with different colors. The figure shows that two clusters always dominate the pooled R&D network: a cluster centered on pharmaceutical companies and a cluster centered on *ICT*-related companies.

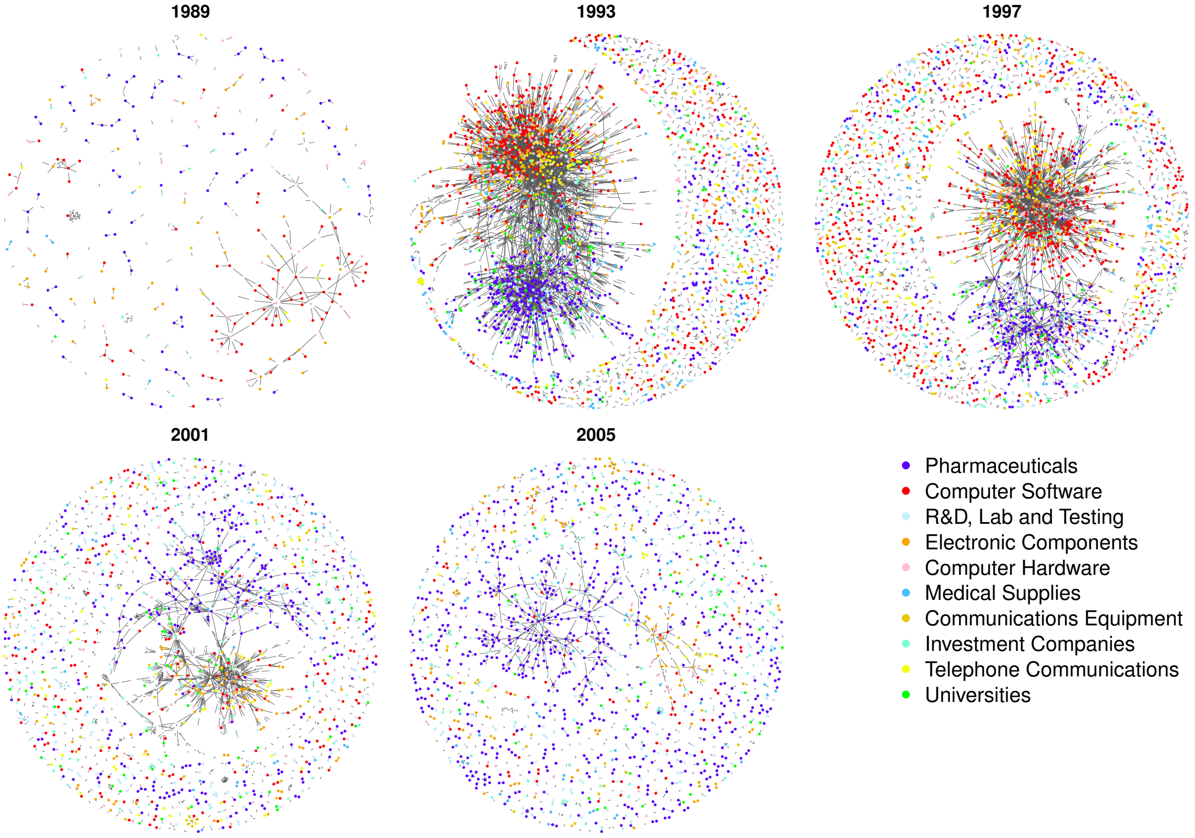


Figure 2.1: Pooled R&D network snapshots in 1989, 1993, 1997, 2001 and 2005. We plot – in different colors – only the ten largest sectors, in order to ease visualization.

Fig. 2.1 denotes the presence of different phases in the evolution of the R&D network. More precisely, the plots suggest the presence of a significant network growth until 1997, and a reversal of this trend in the last periods of our sample. To shed more light on this phenomenon, we report in Table 2.1 the network size, in terms of number of firms taking part in the R&D network – i.e. companies involved in at least one alliance. The observation

[2] The *igraph* library is freely available at `http://igraph.sourceforge.net/`.

period 1986-2009 is divided into six sub-periods of 4 years each and we average the network size within each sub-period. Table 2.1 confirms the presence of a rise-and-fall dynamics in the pooled network. More precisely, the number of companies involved in R&D alliances increases to a peak in the mid-nineties and then shrinks again, both at the pooled and the sectoral level (see Table 2.1). In each sector, the number of firms involved in R&D alliances has a peak in the years 1994-1997. Interestingly, only the Pharmaceutical sector, besides the peak in the period 1994-1997, has an additional peak of slightly larger size in the period 2006-2009. The presence of a peak in the period 1994-1997 is a characteristic of many further network measures considered in this study and leads us to define that period as the "golden age" of R&D networks.

	1986-1989	**1990-1993**	**1994-1997**	**1998-2001**	**2002-2005**	**2006-2009**
Pooled Network	280	2515	4918	2626	2219	1829
Manufacturing Sectors						
Pharmaceuticals (283)	77	645	935	682	825	949
Computer Hardware (357)	51	385	744	202	92	29
Electronic Components (367)	54	328	581	253	222	165
Communications Equipment (366)	17	207	475	181	113	60
Medical Supplies (384)	10	164	280	122	119	123
Laboratory Apparatus (382)	10	139	243	116	94	87
Motor Vehicles (371)	6	108	190	97	85	78
Aircrafts and parts (372)	8	83	136	60	40	26
Inorganic Chemicals (281)	15	108	152	50	45	31
Household Audio-Video (365)	9	110	164	90	65	30
Plastics (282)	11	97	121	44	36	18
Electrical Machinery NEC (369)	2	54	96	26	24	37
Special Machinery (355)	2	33	82	34	17	11
Crude Oil and Gas (131)	3	42	72	62	35	27
Naut./Aeronaut. Navigation (381)	1	49	82	21	16	12
Organic Chemicals (286)	5	44	60	18	23	18
Service Sectors						
Computer Software (737)	69	560	1488	549	284	122
R&D, Lab and Testing (873)	26	477	848	534	596	500
Universities (822)	3	192	374	166	152	83
Telephone Communications (481)	12	184	350	132	82	22
Investment Companies (679)	14	138	298	232	207	125
Professional Equipment Wholesale (504)	4	64	142	26	8	8
Engineer.,Architec.,Survey (871)	2	74	129	62	26	16
Radio and TV Broadcasting (483)	2	26	88	22	7	4
Electric Services (491)	NaN	50	78	38	26	15
Electrical Goods Wholesale (506)	NaN	26	84	19	10	8
Cable and TV Services (484)	NaN	18	78	8	6	3
Motion Picture Production (781)	NaN	15	91	14	4	1
Business Services (738)	1	15	66	37	30	5
Management,Consulting,PR (874)	1	28	96	61	64	28

Table 2.1: Network size of the pooled and the sectoral R&D networks (SIC codes are in brackets). The values are averages within each sub-period. *Note*: missing values refer to sectors with not enough observations.

We show in Fig. 2.2 another visual example of this universal rise-and-fall trend for seven representative industrial sectors. Our plots nicely depict the network snapshots in the years 1989, 1993, 1997, 2001, 2005 and 2009 for the computer software, pharmaceuticals, R&D laboratory and testing, computer hardware, electronic components, communications equipment and universities R&D networks.

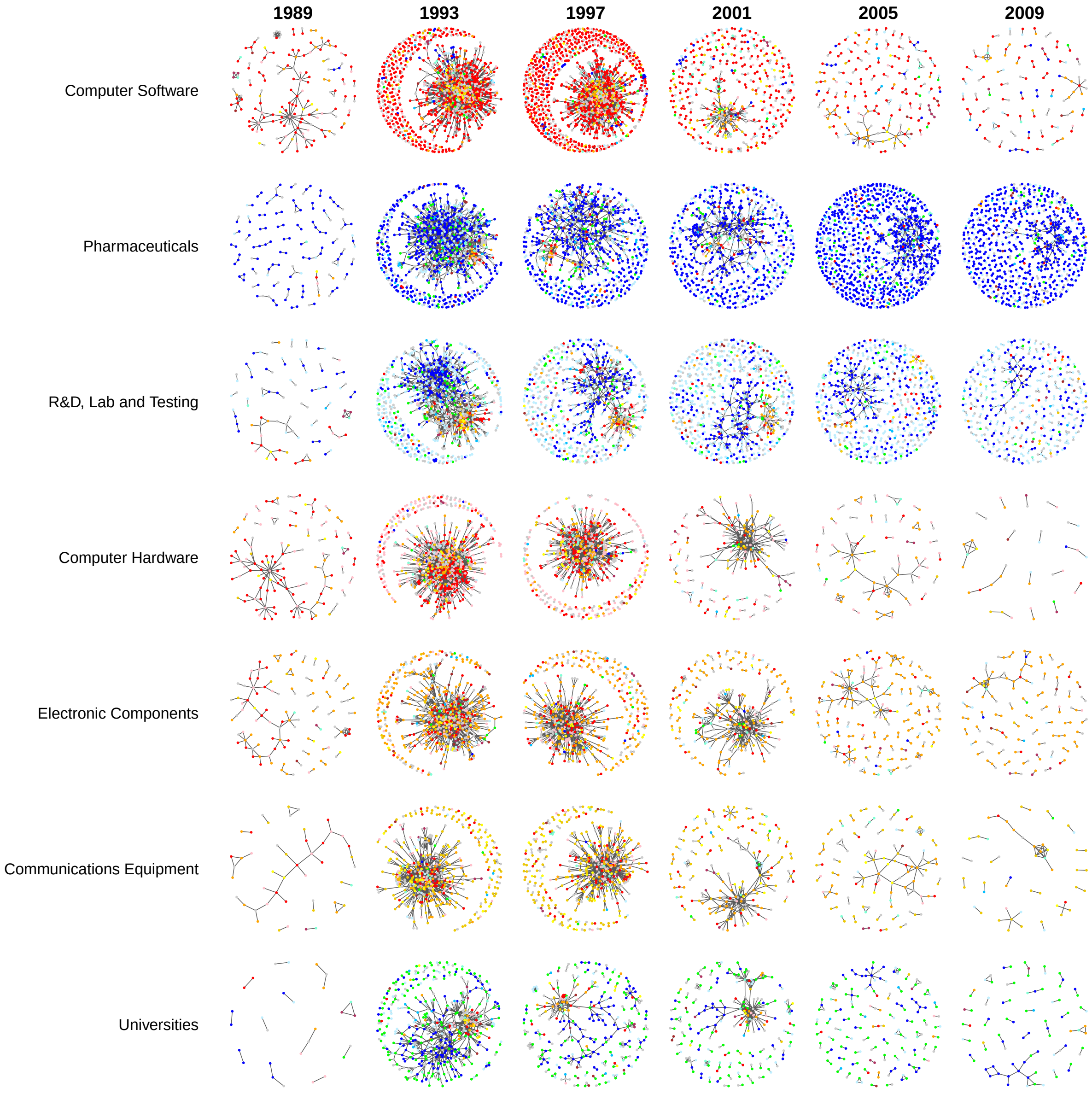


Figure 2.2: Snapshots in 1989, 1993, 1997, 2001, 2005 and 2009 for the seven main sectoral R&D networks. The color legend corresponds to the one reported in Fig. 2.1; firms not belonging to any of the main industrial sectors are depicted in gray.

A deeper investigation shows that the growth in size of the R&D network in the mid-nineties corresponds to a decrease in its density (defined as the number of existing links divided by the number of all possible links in the network). This is shown in Fig. 2.3, where the density of the pooled R&D network, (and its mid-nineties decline), is compared to the network size (and its mid-nineties peak). This means that the expansion of the R&D network was not generated by an increase of the alliances among the firms that were already part of the network. Instead, it was mainly the result of new alliances created by entrant firms. After the "golden age", the shrinking of the network is associated with a decrease in the number of nodes. This fall in the number of firms participating into alliances has however no effect on the density of the network, which remains constant until the end of the observation period (cf. Fig. 2.3).

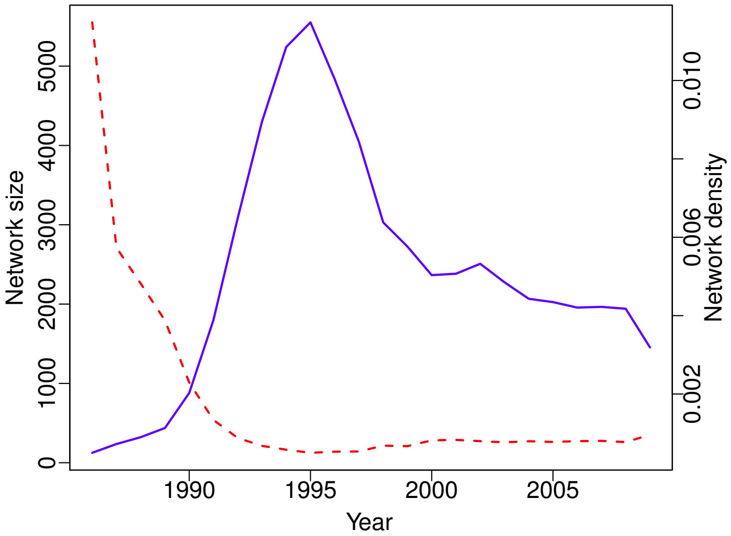


Figure 2.3: Time-evolution of size (solid line, left axis) and density (dashed line, right axis) of the pooled R&D network.

Next, we compute the fraction of nodes belonging to the largest *connected component* of the network. A connected component is defined as a set of nodes which are connected to each other by at least one path (i.e. a sequence of links). We refer to the largest connected component as the *giant component* of the network. The giant component size to the overall network size ratio (or *giant component fraction*) is a rough indicator of the network connectedness. Our results are reported in Table 2.2. This measure has been computed for every year from 1986 to 2009 and then averaged within six sub-periods of 4 years each. Similarly to the network size, the giant component fraction displays a non-monotonic trend at the pooled level, reaching a peak in the mid-nineties and then shrinking again. The emergence of a giant component in the network is of particular interest, as different theoretical works (e.g. Goyal and Joshi, 2003; König *et al.*, 2012) have stressed the importance of the relation between high network connectedness and network efficiency. We also find that the emergence of such non-monotonic dynamics in the giant component is very robust to sectoral disaggregation. Indeed, we observe it in almost all the sub-networks representing the different industrial sectors (see Table 2.2). More precisely, 19 out of the 30 sectoral R&D networks show a giant component peak either in the 1990-1993 or in the 1994-1997 period. The sectors that do not have a peak show a more volatile evolution of their giant component. Among these, only 4 are manufacturing industries (Inorganic Chemicals, Household Audio-Video, Special Machinery, Organic Chemicals), while the other sectors are related to services or sales.

Furthermore, Fig. 2.4 shows the time-evolution of the number of all connected components of the network and of their average size.[3] Both indicators have a peak in the years around 1995 (i.e. the ones corresponding to the 1994-1997 sub-period). This is indicative of the tendency of firms to form more (and larger) connected components until 1995. Afterwards, a fragmentation process takes place. The average size of network components starts to decrease; the number of the components remains stable for two more years, but eventually declines as well (cf. Fig. 2.4). As a result, the large R&D network of the "golden age" period 1994-1997, dominated by a giant component, is replaced by a network with less (and smaller) components. The same results hold for sectoral R&D networks. Fig. 2.1 visualizes this dynamics: the pooled R&D network is characterized by the presence of a giant component that expands until 1997 and subsequently leaves space to a growing periphery of disconnected dyads (pairs of allied firms).

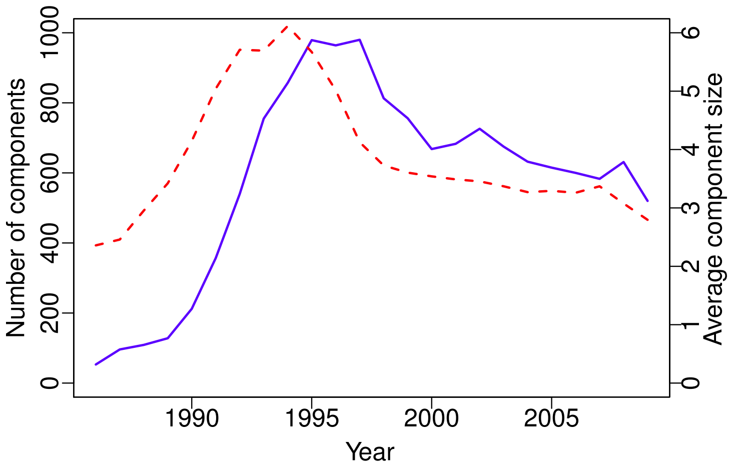


Figure 2.4: Time-evolution of the number of connected components (solid line, left axis) and average size of connected components (dashed line, right axis) in the pooled R&D network.

The above analysis reveals the existence of patterns that are invariant to the scale of aggregation or the sector where they are observed. Namely, both the pooled and sectoral

[3] The distribution of the size components is extremely right skewed and fat-tailed. This is due to the fact that one or few large components co-exist with many disconnected pairs of allied firms. Even though the arithmetic mean is not entirely meaningful or predictive for heavy-tailed distributions, we still report it not only because it is fully computable (we have finite size networks), but also because it gives an idea about the evolution of the component sizes over the period we study. Same remarks apply to the analysis of the average degree that we discuss in Section 2.3.2.

	1986-1989	**1990-1993**	**1994-1997**	**1998-2001**	**2002-2005**	**2006-2009**
Pooled Network	0.10	0.53	0.53	0.33	0.26	0.20
Manufacturing Sectors						
Pharmaceuticals (283)	0.08	0.58	0.68	0.49	0.36	0.32
Computer Hardware (357)	0.27	0.59	0.67	0.51	0.28	0.13
Electronic Components (367)	0.15	0.53	0.61	0.49	0.38	0.13
Communications Equipment (366)	0.18	0.42	0.55	0.25	0.25	0.15
Medical Supplies (384)	0.21	0.04	0.05	0.05	0.06	0.05
Laboratory Apparatus (382)	0.26	0.15	0.13	0.08	0.08	0.07
Motor Vehicles (371)	0.79	0.52	0.39	0.15	0.21	0.10
Aircrafts and parts (372)	0.65	0.47	0.38	0.23	0.20	0.16
Inorganic Chemicals (281)	0.30	0.26	0.17	0.15	0.12	0.29
Household Audio-Video (365)	0.61	0.57	0.61	0.63	0.60	0.28
Plastics (282)	0.23	0.25	0.20	0.23	0.15	0.19
Electrical Machinery NEC (369)	1.00	0.36	0.22	0.20	0.15	0.11
Special Machinery (355)	0.88	0.25	0.13	0.19	0.27	0.26
Crude Oil and Gas (131)	0.67	0.15	0.14	0.10	0.11	0.15
Naut./Aeronaut. Navigation (381)	1.00	0.38	0.26	0.21	0.22	0.24
Organic Chemicals (286)	0.73	0.13	0.17	0.25	0.13	0.22
Service sectors						
Computer Software (737)	0.33	0.54	0.54	0.23	0.11	0.06
R&D, Lab and Testing (873)	0.13	0.19	0.27	0.11	0.10	0.07
Telephone Communications (481)	0.43	0.61	0.58	0.25	0.26	0.28
Universities (822)	0.90	0.17	0.25	0.10	0.08	0.05
Investment Companies (679)	0.21	0.36	0.27	0.23	0.28	0.10
Professional Equipment Wholesale (504)	0.69	0.13	0.16	0.23	0.37	0.28
Engineer.,Architec.,Survey (871)	1.00	0.12	0.15	0.11	0.12	0.20
Motion Picture Production (781)	NaN	0.39	0.24	0.22	0.62	0.50
Management,Consulting,PR (874)	1.00	0.23	0.07	0.09	0.09	0.11
Radio and TV Broadcasting (483)	1.00	0.40	0.17	0.16	0.42	0.61
Cable and TV Services (484)	NaN	0.35	0.16	0.31	0.53	0.75
Business Services (738)	1.00	0.48	0.08	0.11	0.14	0.65
Electrical Goods Wholesale (506)	NaN	0.29	0.12	0.15	0.25	0.34
Electric Services (491)	NaN	0.35	0.11	0.15	0.24	0.21

Table 2.2: Fraction of the giant component of the pooled and the sectoral R&D networks (SIC codes are in brackets). The values are averages within each sub-periods. *Note*: missing values refer to sectors with not enough observations.

R&D networks experience a robust growth in both size and connectedness until 1997. In particular, the years between 1994 and 1997 (the "golden age" of R&D networks), witness not only a higher number of alliances, but also the emergence of a significantly large giant component. This robust growth is then replaced by a decline phase, characterized by both a reduction in the number of alliances and the breaking-up of the network into smaller components. In the next section, we will go into more detail on how these alliances are organized, by studying the degree distributions of the pooled and sectoral R&D networks.

2.3.2 Heterogeneity in alliance behavior

A large part of literature has analyzed the properties of the degree distributions in R&D networks. Empirical studies have shown that degree distributions in R&D networks tend to be highly skewed. Moreover, some studies find exponential distributions (Riccaboni and Pammolli, 2002), while others find power-law distributions (Powell *et al.*, 2005). The presence of a power-law distribution would indicate the existence of an underlying multiplicative growth process (Reed, 2001; Simon, 1955). In the context of R&D networks this means that firms which have many collaborations already attract more new partners than firms with only few collaborations. This idea underlies the "preferential attachment" model by Barabasi and Albert (1999), which predicts the emergence of a power-law degree distribution. However, this model assumes that all firms (even the new entrants) know how many collaborations every other firm in the network has. This may become unrealistic, especially in large networks or situations in which this information is not publicly available. More realistic models assume that firms have only local information about the network. The network formation model introduced by König *et al.* (2014) assumes that firms search for the most central partner in their local neighborhood. Their model generates exponential degree distributions with power-law tails. In the model of Jackson and Rogers (2007), agents also form links locally, which can result in power-law degree distributions as well as exponential degree distributions, depending on various parameters. We extend the existing discussion about the degree distributions in R&D networks by studying their evolution over time and comparing the results between different sectors. Given the small size of many of our networks, we did not test or validate any functional form, but we rather measured the statistical properties of the degree distributions, in order to assess their main features and get insights into the underlying network formation process.

As already mentioned in Section 2.2, we define the degree as the number of partners of a firm, and not the number of alliances. For this reason, we count multiple alliances between the same two firms as one, and we count all the firms participating in the same consortia as distinct partners. Furthermore, like in Section 2.3.1, the whole observation period is divided into six sub-periods lasting 4 years. All the measures we present are computed by aggregating firm degree data relative to the same sub-period. Fig. 2.5 shows the degree distributions of the pooled R&D network in the six analyzed sub-periods. More precisely, given each degree distribution, we report its *complementary cumulative distribution function* $P(x)$, defined as the fraction of nodes having degree greater than or equal to x:

$$P(x) = \int_x^\infty p(x')\mathrm{d}x'. \tag{2.1}$$

where $p(x')$ is the *probability density function*, defining the fraction of nodes in the network with degree x. The complementary cumulative distribution function is more robust than the probability density function against fluctuations due to finite sample sizes (particularly

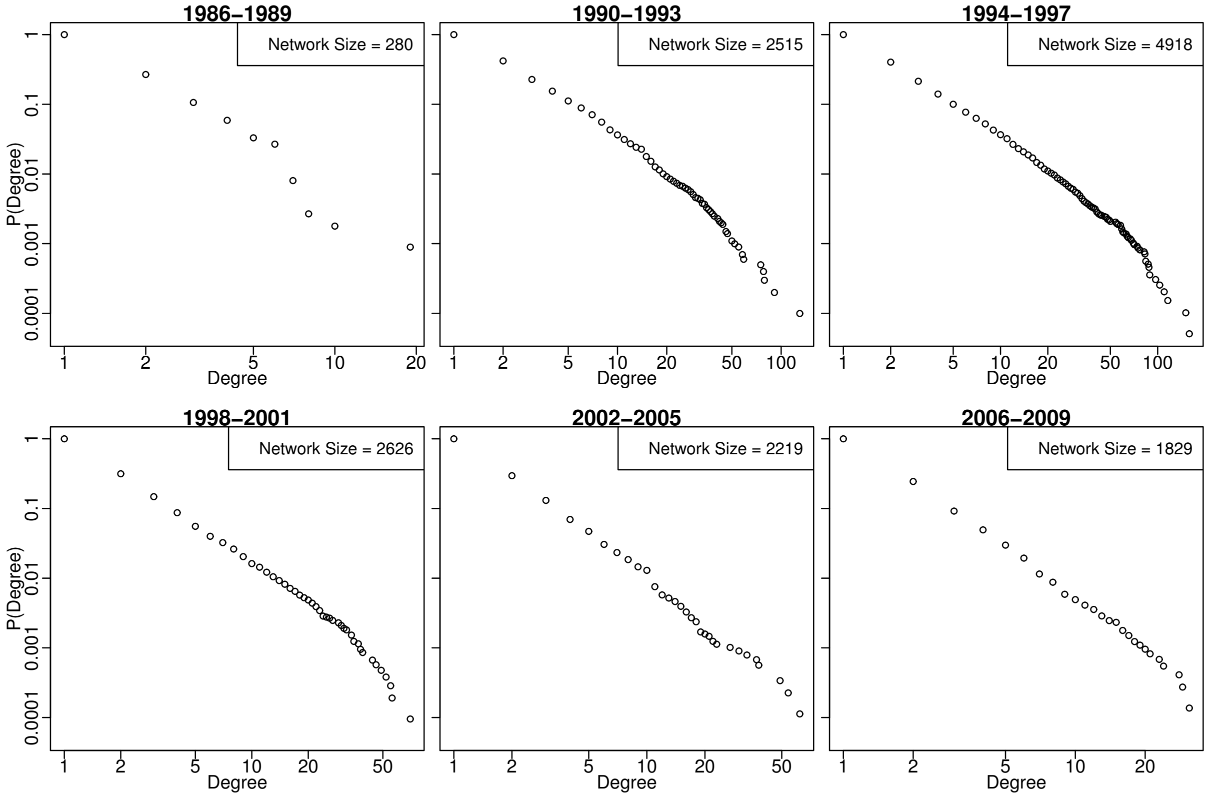


Figure 2.5: Complementary cumulative degree distributions of the pooled R&D network in six sub-periods. *Note:* the insets in the top right corner show the average network size in each of the sub-periods.

in the tail). We find that the degree distribution of the pooled R&D network is very broad and skewed, in all periods. Moreover, the shape of the degree distribution is independent of the network size. For instance, the degree distributions of the pooled R&D network in the "golden age" 1994-1997 (maximum degree $\sim$ 200) has a very similar shape to that of the early period 1986-1989 (maximum degree $\sim$ 20). In addition, most of the sectoral R&D networks (not shown here) exhibit this kind of degree distribution, during the whole observation period.

Table 2.3 shows the first four moments of the degree distribution of the pooled network in each sub-period. In all periods, the degree distribution displays high variance associated with high right-skewness and excess kurtosis. In addition, the p-values of the Kolmogorov-Smirnov test show that the degree distributions of the pooled network are extremely far from the Normal benchmark. Moreover, Table 2.3 shows that all the four moments of the degree distribution increase in the first years of the sample, reaching a peak either in the 1990-1993 or in the 1994-1997 period, and then decrease again. The mean degree has a value of 1.51 partners per firm in the early period 1986-1989; it then exhibits a peak value in 1990-1993 (2.52 partners per firm), which remains almost unchanged in 1994-1997 (2.51 partners per firm), showing that firms have on average more alliance partners in the "golden age" of alliance formation. The average number of partners per firm eventually decreases again, reaching a value of 1.49 in the late period 2006-2009.

As we discussed above, the degree distribution in the pooled R&D network is highly dispersed, as shown by standard deviation values that are always comparable or even larger than the mean values. This holds especially for the 1994-1997 period, when the standard deviation has a peak at 4.98, while the mean value is 2.51 partners per company. Same considerations apply to the evolution of the skewness and kurtosis coefficients over time. In particular, the very high values of the kurtosis coefficient (especially in the period 1994-1997) are indicative of heavy tails in the R&D networks degree distributions, which in turn imply the presence in the networks of "hubs" concentrating a high number of alliances.

	1986-1989	**1990-1993**	**1994-1997**	**1998-2001**	**2002-2005**	**2006-2009**
Mean	1.51	2.52	2.51	1.87	1.70	1.49
SD	1.22	4.30	4.98	2.77	2.11	1.45
Skewness	4.90	9.35	11.28	9.26	10.56	7.92
Kurtosis	47.30	158.40	206.69	133.70	200.25	104.84
KS test p-Value	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-15}$	$< 10^{-15}$

Table 2.3: Degree distribution statistics and p-values of Kolmogorov-Smirnov (KS) test for the pooled R&D network.

The degree distributions of the sectoral R&D networks display patterns that are similar to those of the pooled R&D network.[4] In particular, all sectoral degree distributions are characterized by high variance associated with significant skewness and kurtosis in all sub-periods. We report in Table 2.4 the values of the average degree for the pooled and the sectoral R&D networks in the six sub-periods, clearly confirming such a cross-sector similarity. In all sectoral networks, firms have on average more collaborators during the "golden age" of alliance activity (1994-1997). The only two exceptions are represented by two manufacturing industries, motor vehicles (having a peak in 1986-1989) and organic chemicals (that has a first peak in 1986-1989 and a second one in 1994-1997).

The previous analysis indicates the presence of heavy tails in both the pooled and sectoral degree distributions. In order to get an estimate of the "heaviness" of those tails from a non-parametric point of view, we compute the Hill Estimator (Hill, 1975), a tool commonly used to study the tails of economic data. If n is the number of observations (in our case, the number of nodes in the R&D network) and k is the number of tail observations ($k \leq n$), the inverse of the Hill estimator (HE) is defined as:

$$\hat{h}^{-1} = k^{-1} \sum_{i=1}^{k} \left[\log(x_i) - \log(x_{min}) \right], \tag{2.2}$$

[4] These results are not shown here, but are discussed in Chapters 3 and 4.

	1986-1989	**1990-1993**	**1994-1997**	**1998-2001**	**2002-2005**	**2006-2009**
Pooled Network	1.51	2.52	2.51	1.87	1.70	1.49
Manufacturing Sectors						
Pharmaceuticals (283)	1.22	2.09	2.22	1.82	1.57	1.55
Computer Hardware (357)	1.50	2.10	2.45	2.30	1.55	1.09
Electronic Components (367)	1.32	2.18	2.38	2.15	1.81	1.44
Communications Equipment (366)	1.10	1.82	2.03	1.57	1.48	1.34
Medical Supplies (384)	1.00	1.26	1.31	1.21	1.20	1.16
Laboratory Apparatus (382)	1.00	1.41	1.36	1.24	1.20	1.19
Motor Vehicles (371)	2.31	1.89	1.78	1.40	1.49	1.29
Aircrafts and parts (372)	2.00	2.25	2.00	1.68	1.41	1.40
Inorganic Chemicals (281)	1.28	1.48	1.53	1.23	1.17	1.27
Household Audio-Video (365)	1.44	2.11	2.61	2.32	2.20	1.58
Plastics (282)	1.07	1.54	1.55	1.46	1.29	1.11
Electrical Machinery NEC (369)	1.00	1.45	1.52	1.26	1.11	1.10
Special Machinery (355)	1.00	1.34	1.37	1.24	1.21	1.07
Crude Oil and Gas (131)	1.09	1.70	1.68	1.51	1.28	1.11
Naut./Aeronaut. Navigation (381)	1.33	1.49	1.49	1.23	1.13	1.09
Organic Chemicals (286)	1.26	1.17	1.26	1.14	1.09	1.12
Service Sectors						
Computer Software (737)	1.70	2.16	2.21	1.52	1.27	1.13
R&D, Lab and Testing (873)	1.08	1.68	1.81	1.40	1.43	1.27
Telephone Communications (481)	1.19	2.84	2.53	1.42	1.57	1.28
Universities (822)	1.27	1.66	1.76	1.51	1.35	1.11
Investment Companies (679)	1.04	1.74	1.62	1.53	1.63	1.35
Professional Equipment Wholesale (504)	1.22	1.24	1.42	1.22	1.09	1.00
Engineer.,Architec.,Survey (871)	1.00	1.36	1.40	1.17	1.07	1.09
Motion Picture Production (781)	NaN	1.38	1.36	1.02	1.00	1.00
Management,Consulting,PR (874)	1.00	1.20	1.20	1.19	1.16	1.06
Radio and TV Broadcasting (483)	1.33	1.69	1.31	1.15	1.11	1.11
Cable and TV Services (484)	NaN	1.34	1.51	1.03	1.17	1.00
Business Services (738)	1.00	1.17	1.22	1.15	1.16	1.05
Electrical Goods Wholesale (506)	NaN	1.35	1.34	1.06	1.05	1.07
Electric Services (491)	NaN	1.57	1.38	1.22	1.22	1.25

Table 2.4: Average degree (number of partners) of the pooled and the sectoral R&D networks (SIC codes are in brackets). *Note*: missing values refer to sectors with not enough observations.

where x_{min} represents the beginning of the tail and x_i, $i = 1 \dots k$ are the tail observations, i.e. the degree values such that $x_i \geq x_{min}$. The smaller the HE value, the "heavier" the tail of the degree distribution is. In particular, the degree distributions of most biological, social and economic systems display values of the HE between 2 and 4 (see Clauset *et al.*, 2009). A value of the HE lower than 2 indicates an extremely heavy-tailed distribution ("super heavy-tailedness"). At the other extreme, a value higher than 4 is indicative of degree distributions whose fat-tail property is not very pronounced ("sub heavy-tailedness"). Finally, the theoretical HE value predicted by the preferential-attachment model of Barabasi and Albert (1999) is 3.

Table 2.5 reports the values of the Hill estimator for both the pooled and the sectoral R&D networks in all the time periods. Let us start with the pooled network. The table shows that the HE first decreases, reaching a minimum in the golden-age period 1994-1997 and then increases again. This indicates that the degree of tail-heaviness undergoes a rise-and-fall dynamics similar to the other network measures discussed so far. Moreover, the table shows that in all sub-periods the HE ranges between 2 and 4. This rules out both super and sub heavy-tailedness. However, in all sub-periods but the first and the last one the values of the HE is significantly below 3, and the minimum is achieved in the golden age period 1994-1997 (2.34). This indicates that in those periods the degree distribution of the pooled R&D network cannot be predicted by the preferential-attachment model. In particular, our results show that the tails of the degree distribution of the pooled R&D network are fatter than what will be predicted by that model.

The values of the HE computed on the sectoral R&D networks reveal a rise-and-fall pattern similar to the one detected in the pooled network (see Table 2.5). In particular, most sectors display fatter tails in the periods of higher alliance activity. Moreover, HE values of most manufacturing sectors are comparable to those of the pooled network. In contrast, HE values are in general higher in service sectors. This indicates that the concentration of alliances among few hubs is less marked in this type of sectors.

2.3.3 Degree assortativity

Assortativity is a network measure that identifies correlations between the centrality of a node and the centrality of its neighbors. Assortativity can be computed by using any measure of node centrality (see e.g. Borgatti, 2005, for a survey of centrality measures). However, in this study we use degree correlation, or *average nearest-neighbor connectivity* (Newman, 2002; Pastor-Satorras *et al.*, 2001) as assortativity measure. A network is assortative if it is characterized by a positive correlation across the degrees of linked nodes. This implies that nodes tend to be connected to nodes with similar degree. At the other extreme, dissassortative networks have negative node degree correlation, i.e. nodes tend to be connected to nodes with dissimilar degree. Newman (2003) found that technological networks, such as the Internet, are disassortative while social networks, such as the network of scientific co-authorships, are assortative. However, R&D networks can be assortative or disassortative, depending on the underlying topology of the network. For instance, Ramasco *et al.* (2004) develop models wherein agents establish links with most central actors in the network, and show that such a mechanism gives rise to disassortative networks. However, König *et al.* (2010) show that the same mechanism of search for high centrality can give rise to assortative networks if agents face limitations in the number of collaborations they are able to maintain.

To investigate assortativity-disassortativity in our R&D networks, we use the assortativity mixing coefficient r proposed by Newman (2002). This quantity, as described by Eq. 2.3, is the Pearson correlation coefficient of the degrees at both ends of all links in the network:

	1986-1989	1990-1993	1994-1997	1998-2001	2002-2005	2006-2009
Pooled Network	3.04	2.31	2.34	2.61	2.78	3.05
Manufacturing Sectors						
Pharmaceuticals (283)	5.19	2.91	2.45	2.58	2.89	3.02
Computer Hardware (357)	2.70	2.37	2.22	2.75	2.88	4.59
Electronic Components (367)	3.36	2.43	2.43	2.25	2.59	3.57
Communications Equipment (366)	NaN	2.66	2.50	2.43	2.71	2.65
Medical Supplies (384)	NaN	3.71	3.25	4.50	3.58	3.95
Laboratory Apparatus (382)	NaN	2.69	2.73	3.70	3.22	4.04
Motor Vehicles (371)	3.69	2.18	2.46	2.87	3.72	3.98
Aircrafts and parts (372)	5.07	2.24	2.47	3.77	3.43	3.06
Inorganic Chemicals (281)	3.07	2.31	2.50	3.23	3.71	2.35
Household Audio-Video (365)	3.49	2.48	2.10	2.04	2.09	2.89
Plastics (282)	3.48	3.79	2.34	2.22	3.50	4.36
Electrical Machinery NEC (369)	NaN	3.04	2.89	3.61	4.38	3.29
Special Machinery (355)	NaN	2.89	3.35	3.82	4.44	NaN
Crude Oil and Gas (131)	NaN	3.39	4.08	4.16	6.22	3.59
Naut./Aeronaut. Navigation (381)	NaN	2.53	2.45	4.19	4.10	NaN
Organic Chemicals (286)	3.08	3.86	4.88	4.58	NaN	4.00
Service Sectors						
Computer Software (737)	2.71	2.41	2.30	2.70	3.31	4.24
R&D, Lab and Testing (873)	NaN	2.77	2.69	3.65	3.23	3.60
Telephone Communications (481)	4.63	2.81	2.69	2.94	3.07	3.25
Universities (822)	NaN	2.96	2.72	3.14	3.10	6.01
Investment Companies (679)	NaN	2.86	2.85	2.79	2.85	3.09
Professional Equipment Wholesale (504)	NaN	4.09	3.05	2.60	NaN	NaN
Engineer.,Architec.,Survey (871)	NaN	2.74	2.58	3.14	5.33	NaN
Motion Picture Production (781)	NaN	3.24	3.24	NaN	NaN	NaN
Management,Consulting,PR (874)	NaN	2.91	3.11	3.38	4.43	NaN
Radio and TV Broadcasting (483)	NaN	3.49	3.48	5.17	NaN	NaN
Cable and TV Services (484)	NaN	4.08	3.23	NaN	4.10	NaN
Business Services (738)	NaN	4.59	3.59	4.01	3.59	NaN
Electrical Goods Wholesale (506)	NaN	2.50	2.44	NaN	NaN	NaN
Electric Services (491)	NaN	2.97	3.81	4.01	3.71	NaN

Table 2.5: Hill estimator (HE) for degree distributions of the pooled and the sectoral R&D networks (SIC codes are in brackets). *Note*: missing values refer to sectors with not enough observations.

$$r = \frac{4M^{-1}\sum_i j_i k_i - [M^{-1}\sum_i (j_i + k_i)]^2}{2M^{-1}\sum_i (j_i^2 + k_i^2) - [M^{-1}\sum_i (j_i + k_i)]^2}, \tag{2.3}$$

where j_i, k_i are the degrees of the firms at the ends of the i-th link, with $i = 1, ..., M$. The coefficient r ranges between -1 for a totally disassortative network to 1 for a totally assortative network; a network in which links are formed randomly would exhibit $r = 0$. We compute the assortativity mixing coefficient r on both the pooled and the sectoral R&D sub-networks. We follow the same procedure as in the previous section. The whole observation period is again divided into six sub-periods of 4 years each and all the observations of every firm's degree are taken together within each sub-period. The degree correlation coefficients are then computed for each sub-period. The results are reported in Table 2.6.

	1986-1989	**1990-1993**	**1994-1997**	**1998-2001**	**2002-2005**	**2006-2009**
Pooled Network	0.167	0.110	0.119	0.195	0.170	0.035
Manufacturing Sectors						
Pharmaceuticals (283)	0.005	0.172	0.119	-0.049	-0.047	-0.043
Computer Hardware (357)	-0.188	-0.179	-0.192	-0.133	-0.103	-0.145
Electronic Components (367)	-0.174	-0.151	-0.194	-0.094	0.023	0.267
Communications Equipment (366)	-0.233	-0.149	-0.147	-0.143	-0.077	-0.312
Medical Supplies (384)	NaN	-0.165	-0.155	0.106	-0.184	-0.108
Laboratory Apparatus (382)	NaN	-0.199	-0.134	-0.153	-0.159	0.018
Motor Vehicles (371)	-0.174	-0.309	-0.099	-0.071	-0.023	0.078
Aircrafts and parts (372)	-0.132	0.054	-0.182	0.035	0.019	0.804
Inorganic Chemicals (281)	-0.445	-0.228	-0.243	-0.188	-0.146	-0.239
Household Audio-Video (365)	-0.467	-0.368	-0.306	-0.329	-0.287	-0.342
Plastics (282)	-0.105	-0.249	-0.351	-0.437	-0.265	-0.151
Electrical Machinery NEC (369)	NaN	-0.250	-0.184	-0.283	-0.032	-0.134
Special Machinery (355)	NaN	-0.206	-0.223	-0.153	-0.214	-0.143
Crude Oil and Gas (131)	NaN	0.489	-0.017	0.383	0.255	-0.160
Naut./Aeronaut. Navigation (381)	NaN	-0.297	-0.318	-0.333	-0.217	-0.190
Organic Chemicals (286)	-0.458	-0.242	-0.206	-0.191	-0.190	-0.170
Service Sectors						
Computer Software (737)	-0.103	-0.074	-0.067	-0.029	-0.002	-0.105
R&D, Lab and Testing (873)	-0.024	-0.032	0.011	0.132	0.185	0.025
Telephone Communications (481)	-0.273	-0.178	-0.097	-0.035	-0.036	-0.279
Universities (822)	NaN	-0.133	-0.102	0.026	0.152	0.078
Investment Companies (679)	-0.057	-0.210	-0.193	-0.219	-0.187	-0.182
Professional Equipment Wholesale (504)	NaN	-0.128	-0.066	-0.168	-0.200	NaN
Engineer.,Architec.,Survey (871)	NaN	-0.275	-0.208	-0.130	-0.116	-0.207
Motion Picture Production (781)	NaN	-0.154	-0.081	-0.037	NaN	NaN
Management,Consulting,PR (874)	NaN	-0.288	-0.200	-0.221	-0.177	-0.135
Radio and TV Broadcasting (483)	NaN	-0.537	-0.173	-0.266	-0.250	-0.250
Cable and TV Services (484)	NaN	0.006	-0.101	-0.063	-0.287	NaN
Business Services (738)	NaN	-0.296	-0.247	0.382	0.087	-0.100
Electrical Goods Wholesale (506)	NaN	NaN	-0.305	-0.139	-0.100	-0.143
Electric Services (491)	NaN	-0.007	-0.235	-0.127	-0.107	0.664

Table 2.6: Assortativity mixing coefficient of the pooled and the sectoral R&D networks (SIC codes are in brackets). *Note*: missing values refer to sectors with not enough observations.

The pooled R&D network is assortative, as indicated by the low but positive assortativity mixing coefficient during the whole observation period (see Table 2.6). This means that, on average, high-centrality (low-centrality) firms tend to connect to other high-centrality (low-centrality) firms. Moreover, and differently from the network indicators studied in Sections 2.3.1 and 2.3.2, the assortativity coefficient does not reveal any rise-and-fall dynamics over time.

In contrast to the pooled R&D network, the sectoral R&D networks are disassortative: for most sectors and in most of the analyzed sub-periods, the assortativity coefficient is negative. For instance, when considering the 1990-1993 and the 1994-1997 periods, only 4 sectors out of 30 exhibit a non-negative assortativity coefficient (Pharmaceuticals,

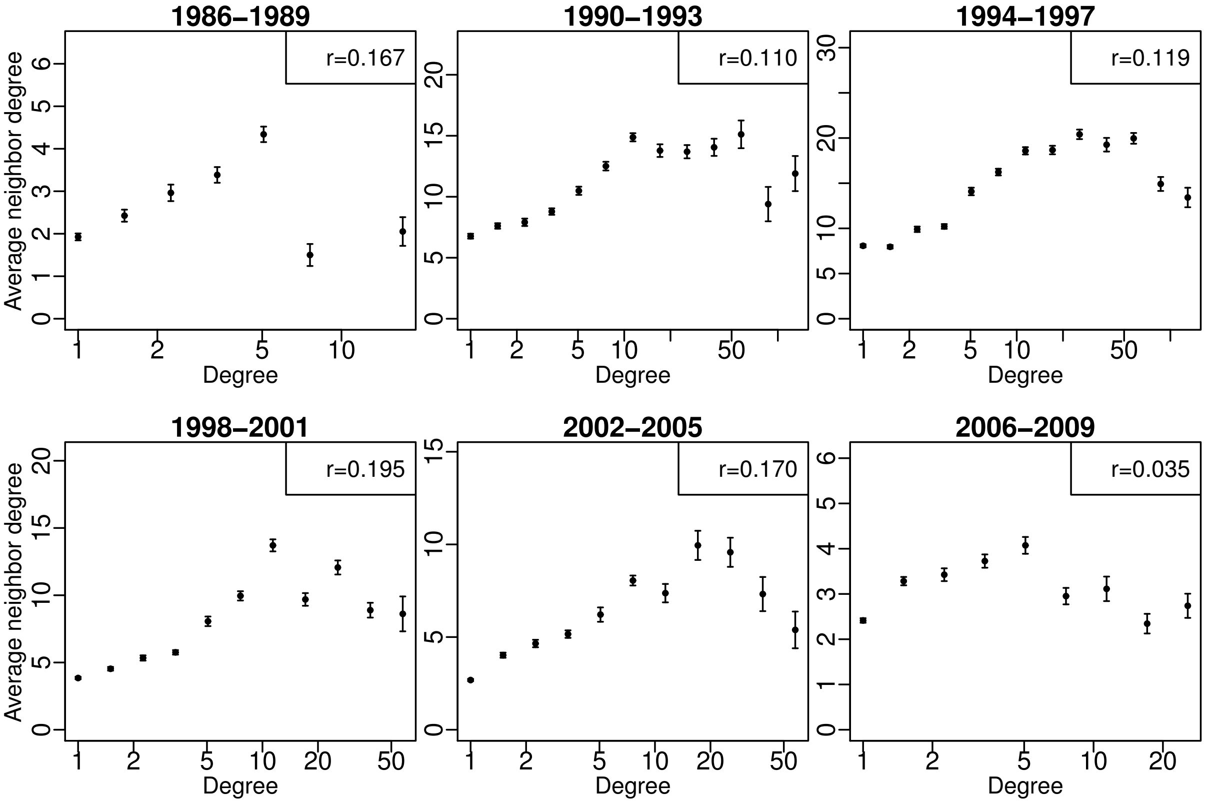


Figure 2.6: Local degree correlations (mean neighbors' degree VS degree) in the pooled R&D network. The error bars represent the standard error of the mean. *Note*: on the top-right corner of each plot we report the corresponding value of the assortativity mixing coefficient in the sub-period under analysis.

R&D-Lab-Testing, Aircrafts and Parts, Cable and TV Services). This indicates that in a sectoral R&D network, i.e. centered around a given industry, low-degree firms increase their tendency to connect to high-degree firms, and vice-versa.

Thus, R&D networks seem to have features of both technological and social networks, as they display both assortativity and disassortativity depending on the scale at which they are studied. To shed more light on the determinants of this phenomenon, we study the "local degree correlations" in the pooled R&D network. More precisely, Fig. 2.6 shows the average neighbors' degree as a function of firms' degree, for the pooled R&D network, and for each of the six sub-periods considered in our analysis.

The plots show that the relation between average neighbors degree and node degree is strongly non linear in all the considered sub-periods. More precisely, node degree predicts quite well average degree of partners until high-degree nodes are taken into account. Then, a sharp decay occurs. This indicates that – when considering the pooled R&D network – firms with low and intermediate degree levels tend to connect with firms having similar degree, whilst high-degree firms display negative degree correlation. Moreover, the position of the maximum of these curves on the x-axis (i.e. the firm's degree) varies during the observation period and is positively correlated to the network size. Such a tipping point in the firm's degree is equal to 5 in the early period 1986-1989 and in the late sub-period 2006-2009, and it ranges between 10 and 20 in the other sub-periods. Interestingly, we find that the inverted U-shaped pattern of the local degree correlation curve holds for the sectoral R&D networks as well. The sharp decay in the local correlation curve is stronger in the sectoral R&D networks than in the pooled one. The above findings indicate that the transition from disassortativity to assortativity is the result of a composition effect due to the presence of a non-linear relationship between the number of alliances of a firm and the one of its partners. In the pooled network sectoral hubs are poorly connected among them (as indicated by the low average degree of their partners). In contrast, firms occupying low and intermediate positions in the sectoral degree distributions tend to form alliances with firms having similar degree in *other* sectors. This does not occur within sectors, where low- and intermediate-degree firms form alliances mainly with the sector hubs.

2.3.4 Small worlds and communities

Similarly to degree heterogeneity (cf. Section 2.3.2) the presence of *small worlds* in R&D networks has been analyzed by a large amount of theoretical and empirical works (see e.g. Cowan and Jonard, 2004, 2009; Fleming *et al.*, 2007; Gulati *et al.*, 2012; Uzzi *et al.*, 2007). A network is a small world if it is characterized by two key features: high local clustering and low average path length (Watts and Strogatz, 1998). Local clustering measures the extent to which the neighbors of a node are in their turn connected among themselves. It is defined as the number of existing links between the neighbors of a focal node, divided by the number of all possible links between these neighbors; the measure is subsequently averaged over all nodes in the network. Average path length is defined as the average of all shortest distances, i.e. the lowest number of links that must be traversed to connect every pair of nodes in the network. In our R&D network representation, the first measure shows the extent to which a company's partners tend to be connected among themselves, while the second measure quantifies how long the average alliance chain from a firm to any other firm in the network is. Small world networks exhibit high clustering and short average path length, combining the qualities of both regular networks (typically characterized by high clustering and high average path length) and random networks (characterized by low clustering and low average path length). Previous empirical works have pointed out that the R&D network structure may follow a rise-and-fall dynamics. More specifically, Gulati *et al.* (2012) show that in the computer industry the excessive formation of ties can lead to the formation of a small world and then to its own decline.

Small world properties in a network are often associated with the presence of community structures (Newman, 2004a), reflecting the tendency of nodes to divide into groups or modules. In a modular network, dense connections and high clustering are observed within each group, with only a few links connecting the different groups (Newman, 2004b). In inter-firm networks, dense groups are shown to facilitate information exchange among similar firms and support trust and cooperative behavior, while bridging ties connecting different groups favor information recombination between distant positions in the knowledge space (e.g. Granovetter, 1973, 1983; Tiwana, 2008).

In this section we analyze both the presence of small worlds and community structures in R&D networks. According to Watts and Strogatz (1998), the small world properties of a network have to be evaluated using a corresponding random network as the baseline. If the examined network is both large and sparse, i.e. $N \gg \langle k \rangle$, where N is the network size and $\langle k \rangle$ is the average degree, the basic requirement for small world is satisfied. Under this assumption, the values of clustering coefficient C and average path length L for the baseline random network will tend to: $C_R = \langle k \rangle / N$ and $L_R = \ln(N)/\ln(\langle k \rangle)$. The small world quotient Q_{SW} we use for our analysis is defined as:

$$Q_{SW} = \frac{(C/C_R)}{(L/L_R)}. \tag{2.4}$$

In our study, the condition of sparse network is always fulfilled for the pooled and the sectoral R&D networks (the average degrees are always smaller than 3, and much smaller than the corresponding network sizes, as reported in Table 2.4). Some of the sectoral R&D networks have relatively small sizes in the first (1986-1989) and in the last (2006-2009) observation periods (as can be seen from Table 2.1), but in these cases they exhibit an even smaller average degree $\langle k \rangle$, still validating the assumption of sparse networks. When computing the observed to random ratios, a small world network will show $C/C_R \gg 1$ and $L/L_R \simeq 1$, which is the case for all the R&D networks we analyze. The results of our computations are listed in Table 2.7. Once again, results are presented for six different sub-periods.

The small world quotient is computed separately for every year during the whole observation period, in both the pooled and the sectoral R&D networks, and then averaged within six sub-periods lasting 4 years each.[5] The evolution of this quotient over time reveals the presence of a rise-and-fall dynamics of the small world properties, in both the pooled and the sectoral R&D networks. The small world quotient rises to a peak in the "golden age" period and then decreases again. Moreover, this feature is common across sectors, generalizing the results of the work by Gulati *et al.* (2012), that was limited to the computer industry. With the exception of 6 sectors out of 30 (Medical Supplies, Universities, Aircrafts and Parts, Business Services, Crude Oil and Gas, Electric Services), the small world

[5] We do not aggregate the observations inside every time period, because the small world quotient is a global network measure, and not an ego-network measure centered around single nodes.

	1986-1989	**1990-1993**	**1994-1997**	**1998-2001**	**2002-2005**	**2006-2009**
Pooled Network	1.410	85.814	154.560	57.085	28.640	5.596
Manufacturing Sectors						
Pharmaceuticals (283)	0.000	23.434	34.030	14.241	5.468	2.628
Computer Hardware (357)	0.129	4.757	16.864	6.397	0.635	0.000
Electronic Components (367)	0.000	7.082	12.414	6.691	5.450	2.290
Communications Equipment (366)	0.000	2.278	5.545	1.283	1.688	0.000
Medical Supplies (384)	NaN	0.000	0.368	0.000	0.000	0.000
Laboratory Apparatus (382)	NaN	0.976	0.534	0.000	0.000	0.933
Motor Vehicles (371)	1.740	2.924	4.134	0.669	1.863	0.840
Aircrafts and parts (372)	1.313	4.319	4.021	1.748	1.323	1.738
Inorganic Chemicals (281)	0.000	0.410	0.000	0.000	0.000	0.000
Household Audio-Video (365)	0.000	1.746	5.316	3.475	1.924	0.000
Plastics (282)	0.000	0.380	0.080	0.000	0.000	0.000
Electrical Machinery NEC (369)	NaN	0.000	0.000	0.000	0.000	0.000
Special Machinery (355)	NaN	0.000	0.000	0.000	0.000	0.000
Crude Oil and Gas (131)	0.000	2.004	0.775	1.269	2.002	0.000
Naut./Aeronaut. Navigation (381)	0.000	0.000	0.000	0.000	0.000	0.000
Organic Chemicals (286)	0.000	0.000	0.000	0.000	0.000	0.000
Service Sectors						
Computer Software (737)	0.769	13.584	33.514	5.242	0.669	0.000
R&D, Lab and Testing (873)	0.000	4.155	12.404	0.864	1.668	0.636
Telephone Communications (481)	0.000	7.521	10.110	1.448	1.222	0.000
Universities (822)	0.000	1.456	4.489	0.863	2.135	0.594
Investment Companies (679)	0.000	0.884	0.452	0.126	0.400	0.576
Professional Equipment Wholesale (504)	0.000	0.594	1.131	0.000	0.000	NaN
Engineer.,Architec.,Survey (871)	NaN	0.450	0.000	0.000	0.000	0.000
Motion Picture Production (781)	NaN	0.000	0.000	0.000	NaN	NaN
Management,Consulting,PR (874)	NaN	0.000	0.000	0.000	0.000	0.000
Radio and TV Broadcasting (483)	0.000	0.000	0.000	0.000	0.000	0.000
Cable and TV Services (484)	NaN	0.320	1.454	0.000	0.000	NaN
Business Services (738)	NaN	0.000	0.000	0.778	0.389	0.000
Electrical Goods Wholesale (506)	NaN	0.000	0.000	0.000	0.000	0.000
Electric Services (491)	NaN	0.429	0.000	0.000	0.594	2.377

Table 2.7: Small world quotient of the pooled and the sectoral R&D networks (SIC codes are in brackets), for the *giant component*. The values are averages within each sub-period. *Note*: missing values refer to sectors with not enough observations.

quotient has a peak either in the 1990-1993 or in the 1994-1997 period. It should also be noticed that five industrial sectors (Motion Picture Production, Management-Consulting-P.R., Electrical Goods Wholesale, Nautical/Aeronautical Navigation, Organic Chemicals) display constant zero values for their small world quotients, meaning that there is no observed clustering in the corresponding networks. The sectors that deviate the most from the non-monotonic small world dynamics are mostly service sectors, which indeed tend to create more inter-sectoral alliances, rather than forming their own intra-sectoral network.

We now want to assess whether such emergence of small world properties in R&D networks is associated with the presence of modular structures. The standard approach to quantify this phenomenon, described by Newman (2004a), is to perform a partition of the network into communities, i.e. assigning a label to every node, in order to maximize the so called *modularity coefficient*. Such indicator of modularity is maximum if the chosen network partition perfectly reflects the positioning of links in the network, with all links occurring within communities and no links occurring between different communities. We do not intend to test several partitions to maximize the modularity coefficient of the network. We rather assume that a community corresponds to an industrial sector. Next, we partition the pooled R&D network by assigning every firm to its sector. Finally, we study the time evolution of the modularity coefficient in the pooled R&D network, computed by considering the sectors as communities. This way, we are able to evaluate the extent to which alliances are concentrated among firms belonging to the same sector. We call the modularity coefficient Q_M and define the *relative connectivity* c_{ij} between two industrial sectors i and j as follows:

$$c_{ij} = e_{ij}/a_{ij}, \tag{2.5}$$

where e_{ij} is the fraction of links in the network connecting any firm belonging to sector i to any firm belonging to sector j. The quantities e_{ij} (and consequently a_{ij} and c_{ij}) can be thought of as elements of a symmetric $n \times n$ matrix, where n is the number of sectors into which the R&D network is partitioned.[6] The row (or column) sums $a_i = \sum_j e_{ij}$ represent the fraction of links (alliances) involving at least one company in sector i. We then define $a_{ij} = a_i a_j$ as the expected fraction of links connecting firms in sector i to firms in sector j in a benchmark network having the same density and sector populations as the real network, but where alliances occur randomly between firms, independently of the sector they belong to. This way, c_{ij} is the ratio between the observed and the expected fraction of alliances connecting a firm in sector i to a firm in sector j. Values of c_{ij} greater than 1 suggest that the alliance probability between a firm in sector i and a firm in sector j is higher than one would expect with a random partner choice. On the contrary, when c_{ij} is smaller than 1, a firm in sector i forms alliances with firms in sector j with a smaller probability than a random partner choice.

We compute all c_{ij} values for the 18 largest industrial sectors analyzed in this Chapter. However, given that the study of intra- and inter-sectoral connectivities is not the main focus of our analysis, we report these results in Appendix A. Next, following Newman (2004a), we define the modularity coefficient Q_M as:

$$Q_M = \sum_i (e_{ii} - a_{ii}^2)/(1 - \sum_i a_{ii}^2), \tag{2.6}$$

where the index i spans all industrial sectors in the R&D network. The coefficient Q_M is equal to 1 in case of a perfect modular network, where alliances occur only intra-community and never inter-community. Likewise, Q_M is equal to -1 for a perfect anti-modular network, having only inter-community links, without any intra-community links.

[6] To make sure that every alliance is counted once in the matrix e_{ij}, every link connecting sectors i and j is split in half between the elements e_{ij} and e_{ji}.

Q_M is equal to zero for a network where links are formed at random. The time evolution of the modularity coefficient Q_M of the pooled R&D network is reported in Table 2.8.

	1986-1989	**1990-1993**	**1994-1997**	**1998-2001**	**2002-2005**	**2006-2009**
Pooled Network	0.237	0.220	0.228	0.220	0.218	0.277

Table 2.8: Modularity coefficients of the pooled R&D network. The values are averages within each sub-period.

The coefficient Q_M ranges between 0.21 and 0.28, indicating the presence of a moderate modularity if compared to other examples of real networks (see Newman and Girvan, 2004). Furthermore, the modularity coefficient exhibits only small changes over the observation period and does not have a peak in accordance with the peak of the small world quotient. The rise-and-fall of the small world structure detected above is thus not associated to any rise-and-fall in the modular structure of the network.

To conclude, our results generalize the previous findings of Gulati *et al.* (2012). The rise-and-fall of small worlds is not a feature limited to few industries, but it is instead a general feature of sectoral R&D networks. Moreover, this property emerges also when alliances are considered independently of the sector to which the firms belong to. However, small worlds are not associated with the presence of a strong community division of the network when industrial sectors are used as communities. The emergence of small worlds might thus have other reasons, which will be further investigated in the next section.

2.3.5 Core-Periphery architectures

Core-periphery networks are dominated by one group of highly inter-connected nodes (the *core* of the network), that have few connections to secondary nodes (the *periphery* of the network). In addition, the peripheral nodes are strongly connected to the core nodes, but poorly inter-connected between each other. Borgatti (2005) points out that such kind of networks are efficient because they can spread information quickly. A generalization of the concept of core-periphery architecture is the one of *nested* networks. A network is nested if the neighbors of a node with degree m are contained in the neighborhoods of all nodes with degree $m' > m$. The difference with core-periphery networks is that graphs with a nested neighborhood structure can feature not only two groups (the *core* and the *periphery*), but several densely connected groups of nodes, with increasing degree. In addition, each group is connected to the group of higher degree nodes. König *et al.* (2012) show that efficient R&D networks (i.e. networks maximizing industry profits) have a nested architecture, when marginal costs of collaborations are high. Interestingly, both core-periphery and nested networks can exhibit short path length and high clustering features that are typical for small worlds. In our case, given the absence of correlation between the emergence of small worlds and modular architectures in the R&D networks, the formation of core-periphery architectures could be the true reason for the emergence of small world properties reported in Section 2.3.4.

To quantify the presence of core-periphery architectures in our R&D networks we employ a slightly modified version of the core-periphery coefficient C_{cp} suggested by Holme (2005).[7] More precisely, we define the core-periphery coefficient C_{cp} of a network G as follows:

$$C_{cp} = \frac{c_c\left[G^{core}\right]/c_c\left[G\right]}{c_c\left[G_R^{core}\right]/c_c\left[G_R\right]}, \tag{2.7}$$

where $c_c[\cdot]$ indicates the closeness centrality of a network[8] and G^{core} is a subgraph[9] of the network G that maximizes this value of closeness centrality. The ratio between the closeness centrality of G^{core} and the closeness centrality of G is then divided by the mean value of the same measure mean value computed on 500 random networks of the same size and density as the network G. The values of the core-periphery coefficients C_{cp} for the pooled and the sectoral R&D networks are shown in Table 2.9. Values are reported – as usual – for the different 6 sub-periods. We do not pool the observations inside each of the 6 selected sub-periods, but we compute the value of the core-periphery coefficient separately for every year and then average over the duration of every sub-period.[10]

We clearly observe a rise-and-fall dynamics for the core-periphery coefficient, in both the pooled and the sectoral R&D networks, with a peak positioned either in the 1990-1993 or in the 1994-1997 period. The presence of core-periphery structures in the "golden age" is a common characteristic across all industrial sectors. One notable exception is the Pharmaceutical sector, whose core-periphery coefficient has a peak in the period 2002-2005. In addition, four small industrial sectors (Management-Consulting-PR, Business Services, Electrical Goods Wholesale and Organic Chemicals) exhibit core-periphery coefficients that are not peaked neither in 1990-1993 nor in the 1994-1997 periods.

The above results confirm that – both at pooled and sectoral level – the small world properties detected in Section 2.3.4 are correlated to the presence of strongly centralized

[7] The difference is that we do not calculate the core-periphery coefficient only on the largest connected component of the network, but we take into account the whole network.

[8] The closeness centrality of a network is defined as the inverse of the sum of all shortest paths between any pair of nodes in the network. The idea behind this measure is to quantify how connected a network is. See Sabidussi (1966) for a more rigorous definition.

[9] There are many ways to divide a network G into subgraphs and then select the subgraph G^{core} with the maximal closeness centrality. Usually, one uses the computationally cheapest algorithm, which is a $k-$core decomposition of the network. G^{core} is then assumed to be the k-shell of the network with maximal closeness centrality. For the sake of brevity, we do not provide here any description of the $k-$core decomposition procedure, See Sabidussi (1966) for a detailed explanation, and Garas *et al.* (2012) for an extension to weighted networks.

[10] Similarly to the small world quotient, the core-periphery coefficient is not an ego-, but a global network measure (see Section 2.3.4).

	1986-1989	**1990-1993**	**1994-1997**	**1998-2001**	**2002-2005**	**2006-2009**
Pooled Network	8.37	23.53	28.51	20.13	18.46	16.34
Manufacturing Sectors						
Pharmaceuticals (283)	0.97	11.41	7.59	12.69	12.88	3.81
Computer Hardware (357)	0.80	5.04	12.38	8.43	2.33	0.12
Electronic Components (367)	1.20	7.17	11.63	6.87	4.39	2.49
Communications Equipment (366)	0.19	3.13	11.05	4.11	2.30	0.77
Medical Supplies (384)	0.30	0.86	3.25	0.98	1.67	1.93
Laboratory Apparatus (382)	0.34	3.95	3.62	2.57	0.04	2.27
Motor Vehicles (371)	0.69	3.25	8.48	2.20	0.91	0.36
Aircrafts and parts (372)	1.87	5.75	9.12	3.18	1.01	1.88
Inorganic Chemicals (281)	0.17	3.60	1.89	0.06	0.07	0.09
Household Audio-Video (365)	0.50	2.98	10.12	5.43	2.96	1.95
Plastics (282)	0.28	2.37	3.80	0.07	0.08	0.16
Electrical Machinery NEC (369)	1.00	2.12	4.24	0.10	0.14	0.09
Special Machinery (355)	0.89	0.93	1.60	0.08	0.24	0.28
Crude Oil and Gas (131)	0.65	3.45	4.20	2.59	1.03	0.11
Naut./Aeronaut. Navigation (381)	1.00	0.78	1.20	0.14	0.19	0.24
Organic Chemicals (286)	0.63	0.11	0.06	0.15	0.13	0.16
Service Sectors						
Computer Software (737)	4.62	8.48	16.38	4.34	5.44	0.03
R&D, Lab and Testing (873)	0.16	4.01	11.36	10.69	5.47	0.49
Telephone Communications (481)	0.39	10.85	14.21	3.12	0.96	0.70
Universities (822)	0.81	0.95	8.47	2.35	1.93	2.35
Investment Companies (679)	0.26	5.75	7.84	7.68	4.91	2.33
Professional Equipment Wholesale (504)	0.58	1.94	2.75	0.15	0.38	0.35
Engineer.,Architec.,Survey (871)	1.00	3.12	2.85	0.05	0.15	0.19
Motion Picture Production (781)	NaN	0.78	1.45	0.30	0.66	0.55
Management,Consulting,PR (874)	1.00	0.19	0.03	0.05	0.05	0.12
Radio and TV Broadcasting (483)	1.00	2.40	2.62	0.13	0.45	0.64
Cable and TV Services (484)	NaN	0.81	4.43	0.36	0.48	0.78
Business Services (738)	1.00	0.46	0.04	1.25	0.72	0.67
Electrical Goods Wholesale (506)	NaN	0.23	0.04	0.18	0.29	0.38
Electric Services (491)	NaN	3.51	1.37	1.42	1.37	2.27

Table 2.9: Core-periphery coefficients of the pooled and the sectoral R&D networks (SIC codes are in brackets). The values are averages within each sub-period. *Note*: missing values refer to sectors with not enough observations.

(core-periphery) architectures. Across sectors, firms show the tendency to organize their R&D collaborations in a core of densely connected companies and a periphery of companies that are linked to the core, but only weakly interconnected among themselves.

Next, we study whether the presence of core-periphery is related to the presence of a more general type of centralized architecture, i.e. nested architectures. In this way we also provide a test to some of the key predictions of the recent theoretical literature on R&D networks. There are several measures quantifying the extent to which a given network's neighborhood structure is nested. In this study, we use the measure generated by an algorithm called *BINMATNEST*[11]. For every analyzed network, the algorithm returns a

[11] The *BINMATNEST* algorithm, proposed by Rodriguez-Girones and Santamaria (2006), uses the unweighted adjacency matrix of the network to compute its nestedness score. The algorithm rearranges

	1986-1989	**1990-1993**	**1994-1997**	**1998-2001**	**2002-2005**	**2006-2009**
Pooled Network	0.977	0.997	0.999	0.997	0.996	0.996
Manufacturing Sectors						
Pharmaceuticals (283)	0.960	0.989	0.996	0.994	0.995	0.996
Computer Hardware (357)	0.960	0.992	0.995	0.984	0.950	0.940
Electronic Components (367)	0.981	0.984	0.993	0.984	0.969	0.926
Communications Equipment (366)	0.943	0.962	0.990	0.973	0.944	0.962
Medical Supplies (384)	0.998	0.944	0.963	0.954	0.946	0.947
Laboratory Apparatus (382)	0.961	0.943	0.965	0.930	0.951	0.924
Motor Vehicles (371)	0.938	0.961	0.964	0.950	0.962	0.946
Aircrafts and parts (372)	0.945	0.942	0.969	0.973	0.953	0.974
Inorganic Chemicals (281)	0.930	0.978	0.951	0.951	0.943	0.977
Household Audio-Video (365)	0.951	0.945	0.981	0.964	0.957	0.963
Plastics (282)	0.977	0.940	0.966	0.951	0.975	0.949
Electrical Machinery NEC (369)	0.939	0.947	0.950	0.962	0.961	0.987
Special Machinery (355)	NaN	0.927	0.940	0.984	0.931	0.953
Crude Oil and Gas (131)	0.939	0.922	0.950	0.945	0.960	0.936
Naut./Aeronaut. Navigation (381)	0.939	0.938	0.948	0.936	0.982	0.998
Organic Chemicals (286)	0.956	0.938	0.961	0.922	0.939	0.956
Service Sectors						
Computer Software (737)	0.981	0.992	0.997	0.985	0.950	0.946
R&D, Lab and Testing (873)	0.961	0.969	0.992	0.986	0.986	0.975
Telephone Communications (481)	0.945	0.954	0.981	0.950	0.947	0.976
Universities (822)	0.961	0.971	0.973	0.952	0.948	0.958
Investment Companies (679)	0.956	0.973	0.979	0.961	0.962	0.940
Professional Equipment Wholesale (504)	0.911	0.930	0.930	0.952	0.921	0.998
Engineer.,Architec.,Survey (871)	NaN	0.924	0.917	0.962	0.957	0.998
Motion Picture Production (781)	NaN	0.935	0.925	0.923	0.937	NaN
Management,Consulting,PR (874)	NaN	0.932	0.933	0.954	0.959	0.939
Radio and TV Broadcasting (483)	0.939	0.941	0.969	0.942	0.967	0.958
Cable and TV Services (484)	NaN	0.930	0.925	0.975	0.973	NaN
Business Services (738)	0.901	0.926	0.952	0.954	0.972	0.998
Electrical Goods Wholesale (506)	NaN	0.951	0.940	0.953	0.935	0.998
Electric Services (491)	0.956	0.918	0.969	0.977	0.957	0.939

Table 2.10: Nestedness coefficients of the pooled and the sectoral R&D networks (SIC codes are in brackets). The values are averaged in six sub-periods. *Note*: missing values refer to sectors with not enough observations.

nestedness score T_n, ranging from 0 (for a totally nested network) to 100 (for a completely random, non-nested network). In order to have a benchmark, the algorithm also builds and analyzes 500 random networks having the same size and density as the considered network. Instead of directly using the value generated by the algorithm, we use a normalized

the adjacency matrix in such a way that all the "ones" (existing links) are concentrated in the top-left side of the matrix, and the "zeroes" (missing links) in the bottom-right side. It then computes the optimal theoretical isocline separating the "ones" from the "zeroes" and counts the number of holes in these regions of the matrix – i.e. how many "zeroes" are in the region of the "ones", and vice-versa. The number of such holes is proportional to the nestedness score computed by the algorithm: the more holes, the higher the nestedness score of the network. *Note:* the lower this score, the more nested the network is (and vice-versa).

nestedness coefficient C'_n, defined as:

$$C'_n = \frac{100 - T_n}{100}, \tag{2.8}$$

where T_n is the nestedness score generated by the *BINMATNEST* algorithm. Our normalized nestedness coefficient C'_n spans thus from 0, for a for a totally non-nested network, to 1, for a totally nested network. We calculate the coefficients C'_n throughout the whole observation period, for the pooled and the sectoral R&D networks, and average the results within six sub-periods lasting 4 years each. Results are shown in Table 2.10.

The values of the nestedness coefficients C'_n we report are extremely close to 1, during the whole observation period, both for the pooled and the sectoral R&D networks. This is surprising, if we compare such values with other studies of nestedness in real networks (e.g. Bascompte *et al.*, 2003). All the values found in our R&D networks are significantly different from the average values of the random networks used as benchmark in the *BINMATNEST* algorithm. Moreover, the nestedness coefficient has a peak during the "golden age" for the pooled R&D network, as well as 9 out of 16 manufacturing sectors and 6 out of 14 service sectors. These results confirm not only that the pooled and the sectoral R&D networks are significantly nested throughout all the observation period, but also that their nestedness tends to increase during the "golden age", in correspondence to the emergence of the small world properties.

2.4 An econometric approach to understand the formation of links

In this section, we investigate the formation of R&D alliances from a microscopic point of view. Differently from the previous section, where the observation unit was an entire (pooled or sectoral) network, now we focus our attention on individual firms. More precisely, we evaluate firms' structural and network features to understand whether and how these features can explain the formation of R&D alliances. Instead of studying the macroscopic properties of a network that is the *result* of R&D alliance formation, we now study how the formation of every single alliance (i.e. every single link in the network) is influenced by a series of firm characteristics.

However, unlike most of the existing studies (e.g. Ahuja, 2000a,b; Powell *et al.*, 1996), we include among the predictors a set of variables that are dependent not only on the firms under examination, but also on the network topology – e.g. the firms' centralities in the network itself.

As our attention is focused on alliance formation, our observation unit is not an individual firm, but a dyad of firms, considered in a given year, irrespectively of whether an alliance actually exists between them. The dependent variable is exactly the formation of an alliance in that given year between the two firms of the dyad: it is a binary variable, equal to 1 if the alliance is formed, equal to 0 otherwise. The independent variables are a set of structural features and network indicators of the two firms, combined in an appropriate fashion. More specifically, we divide such variables into three groups: (a) structural features, (b) network features and (c) potential centrality change, that we describe below in detail.

2.4.1 Model independent variables

Variable group A: structural features. This group contains variables depending on the individual firms or their previous history of alliances, but not on the remaining R&D network as a whole. We have at first two binary variables: belonging of the companies to the same nation (1 if yes, 0 if not) and belonging of the companies to the same industrial sector (1 if yes, 0 if not), evaluated – as we have previously done – by considering their SIC code at a 3-digit level. We then have the number of previous alliances in the dyad, an integer number starting from 0. The last variable in this group is instead a real value, expressing the technological distance between the two companies at the moment of the observation. Such a variable has been extensively used in many studies, to evaluate its effect on alliance formation, or inversely to estimate the effect of alliance formation on firm technological positions. We define a firm's technological position as a $D-$dimensional vector whose components are the share of patents that the firm has in D selected patent classes. The so called technological distance is then the euclidean distances between the two points identified by the coordinates described above in that $D-$dimensional space. In order to evaluate this measure, we use the NBER dataset, listing patent applications in the US patent office classified through the IPC (International Patent Classification) scheme. We select a 1-digit classification level, thus obtaining a total of 8 patent classes. For each time period, we consider all the patents for which the firms applied in the previous five years; if even just one of the firms in a dyad has not applied for any patent in that time window, this will originate a missing observation (this occurs for roughly 60% of our observed dyads). For more details concerning the calculation of this measure, refer to Section 6.2.2; for more details on the concept of a metric $D-$dimensional knowledge space, see Chapters 5 and 6.

Variable group B: network features. This group includes variables describing the position of the two firms in the R&D network. All these variables are related to the focal firms in the dyad, but depend – directly or indirectly – on all the other alliances in the network. In addition, such variables are computed in the year preceding the studied period: for instance, when studying the link formation in 1995, the measures are computed on the R&D network snapshot in 1994. The first of the network variables is the inverse shortest path length between the two firms in the dyad. The shortest path length is defined as the number of links in the network that have to be traversed in order to connect the two firms. This is an integer number ranging from 1 (if the firms are already connected) to infinite (if the firms are isolated or belong to disconnected components). Therefore, the inverse path length is always unequivocally defined and ranges from 0 (for disconnected firms) to 1 (directly connected firms). The second and the third variables are, respectively, the arithmetic mean of the two firms' network centrality, and the difference between the two firms' network centrality in absolute value.

We have tested four different measures of network centrality, namely degree, closeness, betweenness and eigenvector centrality. All of them are highly correlated, as we show in Table 2.12, therefore we decide to employ only one of these four predictors in our analysis. We found that the closeness centrality has the highest predictive power, in terms of Akaike Information Criterion (AIC)[12]. Hence, we will only use this measure, leaving the other centrality measures out of our model. Obviously, using any centrality instead of the closeness would not affect the following results, given their high correlations.

[12] The AIC (Bozdogan, 1987) does not evaluate a model by testing a null hypothesis. Instead, it estimates the goodness of fit of the model and penalizes its complexity. As such, AIC is one of the most used tools for model selection.

Variable group C: potential centrality change. This last group includes variables describing the possible change in a set of network centrality indicators, if the considered dyad forms a link. Among other purposes, this set of variables allows us to test some existing strategic R&D network formation models, in which links are typically formed in order to maximize one (or more) centrality measures. Obviously, we compute such a change in centrality for each dyad irrespective of whether the link is actually formed. We use the network snapshot in the previous year as baseline, and then re-compute all of the centrality measures under examination by adding only that hypothetical link in the network. This way, we can test whether the links that contribute the most to the increase of a given centrality measure are actually formed.

The first variable we use is the average change in closeness centrality of the two firms if the link is formed. That is, we compute the closeness centrality of both firms in the year preceding our observation, and then the same measure for both firms assuming that they form an alliance. The two changes in value for both firm centralities are then averaged.[13] The next variables are directly related to two existing theoretical models. First, we use the change in the eigenvalue of the network connected component to which the firms belong if a given link is formed. This does not have to be confused with the eigenvector centrality, which is a node-centered centrality measure. Here we use instead an aggregate network property, namely the largest eigenvalue of the adjacency matrix of the connected component to which each of the firms of the dyad belongs. We then compute the same measure after an hypothetical link between the two considered firms is formed. The two changes for the eigenvalues of both companies are then averaged.[14] Finally, we use a variable expressing the change of the average harmonic path length of the entire R&D network if a given link is formed. The average harmonic path length is defined as the harmonic mean of all shortest path lengths between all pairs of nodes in the network (considering also disconnected nodes, whose inverse path length is equal to 0). Again, we compute this measure before and after an hypothetical link between the two firms is formed, and the change in value represents our variable.[15]

Control variables. In our regression we use dummy variables for the year in which the dyad is observed – meaning that we use a time-fixed effect model. As we have previously shown, there exists a strong universal trend characterized by a peak of alliance formation in the mid-nineties. We do not want to explain with this model the causes of such rise-

[13] We use the closeness centrality for the same reasons discussed in the previous paragraph (variable group B). The same considerations apply here: any centrality measure could be used without affecting the results.

[14] The use of this variable is inspired by the model in König *et al.* (2012), where the nodes form links maximizing the eigenvalue of the adjacency matrix of the connected component to which they belong. See König *et al.* (2012) for more details.

[15] Such a variable is inspired by the work of Jackson and Wolinsky (1996), where nodes form links to maximize the number of direct and indirect paths connecting them to the other nodes in the network.

and-fall trend. We rather want to understand how the variation of structural and network variables across our firm sample, within each year, can explain the dependent variable, i.e. the formation of a link.

In addition, we control for the number of new alliances formed in total by the dyad – besides the two focal firms themselves – in the observed year. We expect a negative coefficient for this predictor, given that alliances are costly to establish and maintain, as pointed out by several works (Goyal and Joshi, 2003; Goyal and Moraga-Gonzalez, 2001; König *et al.*, 2012). Therefore, the number of new established alliances in any year should be negatively correlated with the formation of one additional alliance. The nomenclature and the meaning of all the variables that we use in our econometric model are reported in Table 2.11.

	Type	Meaning
Dependent variable		
LINK	binary	formation of a link between the two considered firms in the considered year
Controls		
newlinks	positive integer	number of alliances already established by the considered firms in the considered year with other partners
year	binary	dummy variables for the years
Group A		
same_sic	binary	1 if the considered firms have the same SIC code
same_nation	binary	1 if the considered firms are registered in the same nation (source: SDC alliance dataset)
past_alliances	positive integer	number of alliances established between the considered firms in all previous years
tech_distance	positive real	technological distance between the two considered firms in the previous year (measured through patents)
Group B		
inverse_shortest_pl	positive real	inverse of the network path length connecting the two considered firms
closeness_arithm_mean	positive real	arithmetic mean of the closeness centralities of the two considered firms
closeness_difference	positive real	difference (in absolute value) of the closeness centralities of the two considered firms
Group C		
delta_closeness	real	change in the closeness centralities of the two firms (arithmetic mean) if they were to establish an alliance
delta_eigenvalue	real	change in the eigenvalue of the network connected component to which the firms belong (arithmetic mean) if they were to establish an alliance
delta_harmonic_aspl	real	change in the harmonic average path length of the network if the two firms were to establish an alliance

Table 2.11: Nomenclature, type and meaning of all the variables we use in our econometric model.

2.4.2 Results for pooled and sectoral R&D networks

We build seven large panel datasets containing all the above described dyadic variables, for the pooled R&D network and six representative sectoral R&D networks. We start our analysis in 1990 (to avoid missing observations in the early period) and end it in 2009, using a 1-year time window to evaluate the formation of links and all the dyadic measures.

It should be noted that such a procedure generates an enormous amount of data: for each panel, we have $T \cdot N(N-1)$ observations, where T is the length (in years) of the considered time period and N is the number of studied firms. For the sake of computational ease and completeness of data, we restrict our pool to the firms with a significant alliance history, i.e. all firms that have been involved in at least 10 alliance events during the observation period 1990-2009. Yet, all of our panels contain more than 100,000 observation points, even after removing missing observations.

Before proceeding with the actual regression, we show a correlation matrix for all the model variables in Table 2.12, including also the centrality measures we have eventually decided to exclude from our model. This correlation matrix is related to the pooled R&D network, but the sectoral R&D networks panels provide qualitatively unchanged results.

We find that the firm structural features are weakly correlated with both the network variables and the centrality change variables, meaning that the network structure can potentially add a real predictive power to the model, and is not a simple consequence of some static firm attributes, such as their nationality or technological position. Furthermore, we find that the inverse shortest path length between two firms (the first network feature) is significantly correlated with all of the remaining network measures, and with some of the centrality change measures. In other words, the network distance between two firms carries already information about the network centralities of these firms, as well as the potential gain in centrality if the firms would form an alliance – and our regressions will confirm the strength of this predictor for alliance formation.

Finally, as already anticipated, we can observe that most of the network measures are highly correlated among them, meaning that they carry the same information. We find that the closeness centrality mean and difference, besides being the predictors yielding the best AIC score, are those that actually show the lowest correlation with all the remaining network measures. This constitutes one additional reason to use them in our regression.

Finally, we perform seven model regressions on each of the seven panel data sets. Because of the binary nature of our dependent variable, we employ binomial regressions. We choose a complementary log-log link function, which is well suited to very small probability events, such as the formation of an alliance. Indeed, as we already pointed out in Section 2.3.1, the density of any R&D network is very small (ranging between 0.1% and 1%). This is reflected in our panel data sets, which exhibit only 0.16% of successes for the dependent variable (i.e. formation of a link).

The seven models employ, respectively, the structural variables only (model A), the network variables only (model B), the centrality change variables only (model C), then all possible combinations of two variable groups (models AB, AC and BC), ending with a model including all the three variable groups (model ABC). We present the results for the pooled R&D network panel data set in Table 2.13.

	LINK	tech_distance	same_sic	same_nation	inverse_shortest_pl	past_alliances	degree_arithm_mean	betweenness_arithm_mean	closeness_arithm_mean	eigenvectors_arithm_mean	degree_difference	betweenness_difference	closeness_difference	eigenvectors_difference	newlinks	delta_degree	delta_betweenness	delta_closeness	delta_eigenvectors	delta_eigenvalue	delta_harmonic_aspl
LINK	1.00																				
tech_distance	-0.04	1.00																			
same_sic	0.03	-0.26	1.00																		
same_nation	0.01	0.01	0.03	1.00																	
inverse_shortest_pl	0.08	-0.12	0.06	0.05	1.00																
past_alliances	0.00	-0.01	0.00	0.00	0.07	1.00															
degree_arithm_mean	0.08	-0.08	-0.01	0.01	0.59	0.05	1.00														
betweenness_arithm_mean	0.07	-0.06	0.03	0.02	0.40	0.04	0.70	1.00													
closeness_arithm_mean	0.03	-0.05	0.02	0.01	0.63	0.06	0.52	0.33	1.00												
eigenvectors_arithm_mean	0.05	-0.07	-0.05	-0.01	0.39	0.03	0.77	0.41	0.31	1.00											
degree_difference	0.04	-0.06	-0.02	-0.00	0.33	0.03	0.89	0.64	0.36	0.73	1.00										
betweenness_difference	0.05	-0.05	0.03	0.01	0.32	0.04	0.64	0.96	0.28	0.38	0.64	1.00									
closeness_difference	-0.01	-0.00	0.00	-0.05	-0.30	-0.03	0.01	-0.01	0.22	0.05	0.15	0.03	1.00								
eigenvectors_difference	0.02	-0.05	-0.06	-0.02	0.26	0.03	0.70	0.38	0.27	0.95	0.75	0.36	0.10	1.00							
newlinks	0.16	-0.06	-0.02	-0.00	0.27	0.02	0.46	0.39	0.23	0.31	0.42	0.37	0.01	0.28	1.00						
delta_degree	-0.10	0.07	-0.05	-0.02	-0.47	-0.01	-0.19	-0.14	-0.08	-0.16	-0.08	-0.10	0.05	-0.05	-0.09	1.00					
delta_betweenness	-0.00	0.08	-0.05	0.02	0.23	0.06	0.36	0.31	0.28	0.20	0.23	0.24	-0.09	0.23	0.17	0.02	1.00				
delta_closeness	-0.02	0.05	-0.02	-0.04	-0.48	-0.05	-0.26	-0.17	-0.21	-0.13	-0.12	-0.13	0.55	-0.08	-0.13	0.07	-0.22	1.00			
delta_eigenvectors	0.03	-0.05	-0.02	-0.02	0.14	0.01	0.29	0.14	0.17	0.60	0.28	0.13	0.11	0.61	0.12	-0.08	0.07	0.01	1.00		
delta_eigenvalue	-0.01	0.01	-0.01	-0.04	-0.37	-0.03	-0.02	-0.01	-0.10	0.02	0.14	0.04	0.60	0.06	-0.01	0.05	-0.10	0.33	-0.01	1.00	
delta_harmonic_aspl	-0.01	0.04	-0.01	0.02	-0.16	-0.03	-0.24	-0.13	-0.38	-0.17	-0.20	-0.12	-0.23	-0.17	-0.12	0.02	-0.22	0.40	-0.16	-0.17	1.00

Table 2.12: Pearson correlation coefficients for all pairs of variables used in our econometric model (plus the three additional centrality measures that we have eventually discarded, i.e. degree, betweenness and eigenvector centralities).

By comparing the AIC scores, we find that all seven models exhibit similar goodness of fit. However, the variable group A (structural firm features) has a slightly higher predictive power than the group B (network firm features), which in its turn has a higher predictive power than the group C (centrality change variables). It should be noted that the structural variables alone (model A) perform better than the network variables and the centrality change variables combined (model BC). Nevertheless, the model exhibiting the best AIC score is the complete model ABC, including all of the three variable groups. This means that the alliance formation is optimally explained by a combination of structural and network firm variables; however, firm structural variables alone have a slightly better explanatory power than network variables alone.

We find that most predictors in all three groups are significant; their effect is stable in sign and magnitude across all models. In particular, the effect of the binary variables "same nation" and "same SIC" is positive – as expected, geographical and sectoral proximities positively affect alliance formation. Likewise, the technological distance has a negative effect on alliance formation, showing that firms with closer patenting activities are more likely to establish new alliances. The variable "new links", contrary to what we expected, shows a positive effect on the dependent variable, meaning that firm dyads that have

Model	A	B	C	AB	AC	BC	ABC
(Intercept)	$\mathbf{-5.638}^{***}$	$\mathbf{-7.094}^{***}$	$\mathbf{-4.931}^{***}$	$\mathbf{-6.023}^{***}$	$\mathbf{-4.535}^{***}$	$\mathbf{-5.613}^{***}$	$\mathbf{-4.867}^{***}$
	(0.122)	(0.126)	(0.175)	(0.137)	(0.185)	(0.225)	(0.236)
newlinks	$\mathbf{0.214}^{***}$	$\mathbf{0.190}^{***}$	$\mathbf{0.227}^{***}$	$\mathbf{0.189}^{***}$	$\mathbf{0.211}^{***}$	$\mathbf{0.191}^{***}$	$\mathbf{0.190}^{***}$
	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)	(0.003)
same_nation	$\mathbf{0.594}^{***}$			$\mathbf{0.534}^{***}$	$\mathbf{0.594}^{***}$		$\mathbf{0.536}^{***}$
	(0.039)			(0.040)	(0.039)		(0.040)
same_sic	$\mathbf{1.044}^{***}$			$\mathbf{0.968}^{***}$	$\mathbf{1.035}^{***}$		$\mathbf{0.961}^{***}$
	(0.045)			(0.045)	(0.045)		(0.045)
past_alliances	0.138			−0.009	0.097		−0.011
	(0.096)			(0.118)	(0.103)		(0.119)
tech_distance	$\mathbf{-2.656}^{***}$			$\mathbf{-2.279}^{***}$	$\mathbf{-2.599}^{***}$		$\mathbf{-2.276}^{***}$
	(0.078)			(0.080)	(0.078)		(0.080)
inverse_shortest_pl		$\mathbf{3.232}^{***}$		$\mathbf{2.118}^{***}$		$\mathbf{3.249}^{***}$	$\mathbf{2.159}^{***}$
		(0.083)		(0.089)		(0.083)	(0.089)
closeness_arithm_mean		0.151		0.610		$\mathbf{-3.431}^{***}$	$\mathbf{-2.781}^{**}$
		(0.728)		(0.739)		(0.894)	(0.926)
closeness_difference		$\mathbf{4.028}^{***}$		$\mathbf{3.282}^{***}$		$\mathbf{5.227}^{***}$	$\mathbf{4.120}^{***}$
		(0.431)		(0.435)		(0.801)	(0.791)
delta_closeness			$\mathbf{-9.069}^{***}$		$\mathbf{-4.835}^{***}$	$\mathbf{-6.703}^{***}$	$\mathbf{-4.710}^{***}$
			(0.822)		(0.799)	(1.053)	(1.053)
delta_eigenvalue			$\mathbf{0.059}^{***}$		$0.024^{\cdot}$	$\mathbf{0.059}^{***}$	$\mathbf{0.036}^{**}$
			(0.012)		(0.012)	(0.013)	(0.013)
delta_harmonic_aspl			−0.097		$\mathbf{-0.330}^{***}$	$\mathbf{-0.263}^{**}$	$\mathbf{-0.380}^{***}$
			(0.094)		(0.088)	(0.100)	(0.099)
AIC	29105.534	30527.208	31783.977	28468.056	29000.683	30474.512	28433.889
BIC	29415.006	30824.301	32081.070	28814.665	29347.291	30808.741	28817.633
Log Likelihood	-14527.767	-15239.604	-15867.989	-14206.028	-14472.341	-15210.256	-14185.944
Deviance	29055.534	30479.208	31735.977	28412.056	28944.683	30420.512	28371.889
Num. obs.	1756561	1756561	1756561	1756561	1756561	1756561	1756561

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 2.13: Results of the regressions for our econometric model on the pooled R&D network. The coefficients with p-value smaller than 0.01 are reported in bold character.

formed more alliances in a given year (excluding the dyad itself) are more likely to partner with each other as well and close the dyad. In other words, our data do not show limitation effects in the number of alliances that firms are able to establish in a given year. Interestingly, the number of past alliances is never significant for the formation of a new alliance between two firms.

As for the network variables, we find that the inverse shortest path length in a dyad always has a strong positive effect on alliance formation. This means that the establishment of a new alliance is more likely between firms that are linked by a path in the network (including the case in which they already have an alliance). The mean closeness centrality of the dyad is significant only when considered together with the centrality change measures, and its sign is negative, meaning that dyads with an overall low closeness centrality are more likely to form a new alliance. The closeness difference of the dyad is instead always significant, with a positive sign, meaning that dyads with a larger closeness centrality disparity are more likely to form a new alliance. A firm dyad with low centrality mean and high centrality absolute difference corresponds to a dyad where one of the firms has low centrality and the other one has high centrality; therefore, firms with a larger centrality disparity contribute more to alliance formation.

The centrality change measures are surprisingly stable and significant, especially if we consider them together with the firm network variables. The closeness centrality change is always significant and – interestingly – has a negative effect on the alliance formation, meaning that the links causing a larger mean increase in the partners' centralities are actually less likely to be formed. On the contrary, the last two variables, expressing the change in aggregate network centralities – as opposed to individual firm centralities – show an overall positive effect on the dependent variable. In particular, the change in the eigenvalue of the firms' connected component adjacency matrix always exhibits a positive coefficient, with a $p-$value not greater than 10%, meaning that the links providing the highest increase in this eigenvalue are the most likely to be formed. Likewise, the change in the harmonic average path length of the R&D network is significant when considered together with the other firm structural and/or network features, and exhibits a negative coefficient, meaning that links causing a higher decrease in the average path length of the whole R&D network are more likely to be formed.

The fact that link creation is favored when it increases aggregate network cohesiveness measures, instead of individual node centralities, is a surprising finding. However, it is not enough to infer that firms intentionally form alliances that increase their centrality the least, preferring some aggregate network benefit; it just means that the complex interdependencies between the firm strategy and the network growth give rise to this kind of pattern in the data. Indeed, we believe that firms do form alliances trying to increase their own utility, which clearly does not depend solely on network indicators. Only an agent based model can give us better insights and reproduce an environment where firms strive to maximize their individual utility and yet the resulting link formation leads to an increase of the aggregated network cohesiveness.

We then present in Table 2.14 a summary of the results for the pooled R&D network, together with the seven representative sectors we have selected. For all columns, the coefficients are related to the complete model ABC, i.e. the one including all of the three variable groups. The complete results for the other model variants on these seven sectoral R&D networks are reported in Appendix A.

We find that most of our selected predictors exhibit a robust behavior across sectors. More precisely, the following variables display significant and stable coefficients in all panels: "new links" (positive effect), "same SIC" (positive effect), "technological distance" (negative effect) and "inverse shortest path length" (positive effect). The variable "same nation" is always significant and with a positive effect on the alliance formation, with the only exception of the Medical Supplies sector (where it does not have a significant effect). The variable "past alliances", similarly to the pooled R&D network, is generally not

Sector (SIC code)	**Pooled**	737	367	357	366	283	873	384
(Intercept)	**−4.867*****	**−4.647*****	**−4.852*****	**−4.588*****	**−4.636*****	**−4.467*****	**−4.537*****	−2.634*
	(0.236)	(0.234)	(0.288)	(0.234)	(0.281)	(0.341)	(0.410)	(1.289)
newlinks	**0.190*****	**0.175*****	**0.212*****	**0.199*****	**0.237*****	**0.481*****	**0.307*****	**0.941*****
	(0.003)	(0.004)	(0.005)	(0.004)	(0.006)	(0.010)	(0.009)	(0.077)
same_nation	**0.536*****	**0.496*****	**0.761*****	**0.648*****	**0.695*****	0.169*	**0.206****	−0.004
	(0.040)	(0.056)	(0.055)	(0.056)	(0.059)	(0.068)	(0.067)	(0.250)
same_sic	**0.961*****	**0.349*****	**0.348*****	**0.329*****	**0.448*****	**0.454*****	**0.700*****	**0.752****
	(0.045)	(0.067)	(0.068)	(0.067)	(0.075)	(0.079)	(0.076)	(0.282)
past_alliances	−0.011	−0.241*	−0.020	**0.116****	−0.157*	0.128	0.060	0.222
	(0.119)	(0.112)	(0.064)	(0.037)	(0.070)	(0.101)	(0.088)	(0.249)
tech_distance	**−2.276*****	**−1.630*****	**−1.234*****	**−1.298*****	**−1.077*****	**−2.095*****	**−2.576*****	**−2.721*****
	(0.080)	(0.106)	(0.120)	(0.110)	(0.122)	(0.139)	(0.139)	(0.543)
inverse_shortest_pl	**2.159*****	**1.833*****	**1.523*****	**1.569*****	**1.315*****	**2.218*****	**1.721*****	**2.089*****
	(0.089)	(0.120)	(0.125)	(0.127)	(0.133)	(0.146)	(0.137)	(0.393)
closeness_arithm_mean	**−2.781****	0.149	**0.333****	−0.042	**0.203****	**−0.674*****	0.326	−0.031
	(0.926)	(0.122)	(0.123)	(0.099)	(0.073)	(0.179)	(0.321)	(0.027)
closeness_difference	**4.120*****	**0.439****	−0.157	0.246$^{\cdot}$	−0.130	**0.549*****	0.002	0.032
	(0.791)	(0.170)	(0.146)	(0.130)	(0.105)	(0.162)	(0.198)	(0.021)
delta_closeness	**−4.710*****	**−0.847*****	−0.332$^{\cdot}$	**−0.733*****	**−0.329****	−0.435*	−0.098	−0.099*
	(1.053)	(0.240)	(0.181)	(0.189)	(0.125)	(0.204)	(0.234)	(0.045)
delta_eigenvalue	**0.036****	**0.109*****	**0.102*****	**0.105*****	**0.132*****	0.036	0.063*	−0.088
	(0.013)	(0.016)	(0.019)	(0.022)	(0.019)	(0.022)	(0.025)	(0.102)
delta_harmonic_aspl	**−0.380*****	**−1.047*****	**−0.813*****	**−1.170*****	**−0.747*****	−0.200*	−0.139	−0.051
	(0.099)	(0.239)	(0.164)	(0.220)	(0.179)	(0.095)	(0.095)	(0.234)
AIC	28433.889	11900.390	11304.548	11184.447	9061.810	8935.185	9344.910	605.090
BIC	28817.633	12219.859	11619.243	11497.831	9361.489	9265.528	9673.273	832.585
Log Likelihood	-14185.944	-5919.195	-5621.274	-5561.223	-4499.905	-4436.593	-4641.455	-271.545
Deviance	28371.889	11838.390	11242.548	11122.447	8999.810	8873.185	9282.910	543.090
Num. obs.	1756561	220896	189365	181529	116668	313709	294304	11368

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table 2.14: Results of the regressions of our complete econometric model (ABC, including all variables) on the pooled and the seven main sectoral R&D networks, namely Pharmaceuticals (SIC code 283), Computer Hardware (SIC code 357), Communications Equipment (SIC code 366), Electronic Components (SIC code 367), Medical Supplies (SIC code 384), Computer Software (SIC code 737) and R&D, Laboratory and Testing (SIC code 873). The coefficients with p-value smaller than 0.01 are reported in bold character.

significant, except in the Computer Software sector (negative effect), the Communications Equipment sector (negative effect) and the Computer Hardware sector (positive effect).

The network centrality variables (i.e. closeness mean and absolute difference) show a large variance across sectors and they do not seem to have a significant effect on alliance formation in the sectoral R&D networks. On the contrary, the centrality change measures show a fairly robust behavior across sectors, and replicate the trend that we have observed in the pooled R&D network. Namely, the change in individual firm centralities is either negatively affecting the alliance formation or not significant; the change in the eigenvalue of the firms' connected component is either positively affecting the alliance formation or not significant; and the change in the harmonic average path length of the network is either negatively affecting the alliance formation or not significant.

In conclusion, while the network variables (group B) exhibit a large sectoral variance, the structural firm variables (group A) and the centrality change variables (group C) are fairly robust across sectors and reveal a set of tendencies that can be summarized as follows. In R&D networks, alliances are more likely to be established if:

- the partners belong to the same country and industrial sector, and they have a small technological distance;
- the partners have already engaged in many alliances with other distinct firms;
- the partners are already – directly or indirectly – connected by a path in the R&D network;
- the partners have a low mean centrality and a high centrality difference, i.e. one of the two has a high centrality and the other one has a low centrality in the R&D network (this feature is significant in the pooled R&D network, but not robust across sectors);
- the formation of the considered link leads to a small increase in the individual firm centralities, but a large increase in a set of aggregate network centrality indicators.

2.5 Discussion

Four main implications arise from the evidence discussed in the present chapter. *First*, our results provide strong support to the claim that several properties of R&D networks are robust across several manufacturing and service sectors. These properties are invariant across different scales of aggregation as well. In other words, they are the same if one considers the R&D alliances irrespectively of the sectors to which the firms belong (pooled network), or if one considers only alliances centered on a sector (sectoral networks). These properties do not only relate to basic network characteristics like size, density, degree distributions. They also involve more complex features such as the presence of small worlds and core-periphery architectures, and the microscopic rules determining the formation of the alliances themselves. For instance, alliance preferences between firms with small geographical, sectoral, technological and network distances are stable and robust across sectors. From an empirical perspective, our results thus generalize previous findings in the literature, that were limited to the analysis of few sectors. From a theoretical perspective, the fact that many properties of the network hold irrespectively of the sector and of the scale of aggregation opens up the fascinating possibility that the same universal mechanism can be responsible for the emergence of those features.

In this respect, our results also show that such a mechanism is probably different and more sophisticated than the preferential attachment described by Barabasi and Albert (1999).

This is because the characteristics of the degree distribution observed in our R&D networks (cf. Section 2.3.2) can be hardly reconciled with the predictions of that model. Nevertheless, our results also show that not *all* properties of the network are invariant across different scales of aggregation. Sectoral networks are disassortative, i.e. characterized by a negative correlation across node degrees, whereas the pooled network is assortative. This transition from disassortative to assortative networks is a fresh new stylized fact that should be taken into account in the theoretical explanations of R&D networks. It is important to remark that the contrast between disassortative and assortative networks has been so far stressed in relation to networks belonging to different domains (e.g. technological vs. social networks, cf. Newman, 2003). Our results suggest instead that the same type of network (network of R&D alliances) can be disassortative or assortative depending on the scale at which it is observed (i.e. taking into account the sectoral characteristics of the partners or not). This instability of degree-degree correlations in R&D networks is reflected in our findings at the microscopic level. While in the pooled R&D network the firms contributing the most to the alliance formation exhibit a strong centrality disparity, in sectoral R&D networks this tendency disappears to be replaced by sectoral specific behaviors.

Second, the result that both the pooled and sectoral networks are organized into core-periphery architectures – nested structures in particular – militates in favor of the predictions of the recent theoretical literature on R&D networks (e.g. Goyal and Joshi, 2003; Westbrock, 2010), and more precisely of the knowledge-recombination model of König *et al.* (2012). In this model, the efficient network structure is shown to critically depend on the marginal cost of R&D collaborations. In case of relatively costly partnerships, the resulting efficient R&D network exhibits a strongly nested neighborhood structure, as we observe empirically. Moreover, the presence of core-periphery architectures is also able to explain two network properties that received a lot of attention in the literature, namely the emergence of fat-tailed degree distributions and of small worlds. These properties are indeed the result of the organization of the R&D networks into core-periphery structures and cannot be instead related to other types of network characteristics (e.g. the presence of communities for small worlds, as we show in Section 2.3.4).

These findings are further supported by our econometric approach, which shows – interestingly – that alliances are more likely to be observed if they maximize the change in some aggregate network measures, i.e. the eigenvalue of the connected component to which the firms belong (König *et al.*, 2012) or the harmonic average path length of the network (Jackson and Wolinsky, 1996), and not the increase of the single firm centralities. The network topology resulting from this behavior, again, is compatible with the observed nested architectures. Most likely, this does not mean that firms are not concerned with the improvement of their own network centrality when establishing new alliances. We argue instead that firms do try to maximize their expected return, but this may depend on both network-related and network-unrelated factors. The complex interdependencies between firm decisions and the actual alliance formation give rise to a network growth process where the newly formed links tend to maximize some aggregate network indicators rather than individual firm centralities. The result is the observed coefficient sign in our econometric model. Similarly, the data do not show any evidence of another theoretical aspect, i.e. costly R&D alliances: firm dyads engaged in other distinct alliances are more likely to form an alliance themselves, thus not showing any limitation effect in the number of newly established R&D alliances. Again, this might be an effect of the complex interdependencies above mentioned; only the use of an agent based model – that we investigate in the next chapters – will be able to give us further insights.

Third, previous network structures, along with potential network structure changes, matter in the alliance formation, as testified by the shape of the degree distribution in R&D network, as well as by our econometric approach. Even though the network-unrelated variables alone have a slightly better predictive power than network-related variables alone, we have shown that a model including both types of variables has the highest possible goodness of fit when explaining the formation of R&D alliances. In addition, the analysis of the predictor coefficients allows us to identify an additional set of invariant and sectoral-robust features at the microscopic level. Namely, alliances are more likely to be established if the potential partners belong to the same country and sector, and exhibit a small technological distance; if they have already engaged in many alliances with other distinct firms; if they are already – directly or indirectly – connected by a path in the R&D network; if the formation of the considered link leads to a small increase in the individual firm centralities, but a large increase in a set of aggregate network centrality indicators, as already discussed.

Fourth, our evidence indicates that the last three decades have witnessed a rise and fall of R&D networks. The foregoing rise-and-fall dynamics was previously emphasized in relation to the presence of small worlds in the computer industry (Gulati *et al.*, 2012). We show that it is instead a general property of the R&D network dynamics (both sectoral and pooled ones). In addition, it concerns even more complex network properties (presence of core-periphery and nested architectures).[16] Our results also show that the rise and fall of R&D networks was mainly driven by the entry and exit of firms participating into alliances rather than by the more or less intense activity of the incumbents (see Section 2.3.1). Moreover, during the growing phase, R&D alliances gave rise to network components of large size displaying the complex features discussed above. In the descending phase, the number of firms participating into alliances plummeted and the networks broke up into several components of small size.

Overall, the above facts suggest that theoretical explanations of the dynamics of the net-

[16] Indeed, the only exception to this general dynamics is represented by the assortativity (resp. disassortativity for sectors), that does not display any particular pattern over time.

work should account for a significant role of the entry/exit of firms. In addition, they should be able to explain the ability of the network to self-organize into components having complex characteristics and the eventual breaking-up of them. Finally, as it is argued in Gulati *et al.* (2012), the rise and fall of R&D networks could be the sheer outcome of the knowledge recombination process associated with alliances embedded into a network. Indeed, the possibility of knowledge recombination fuels the growth of the network, either by combining heterogeneous knowledge bases (e.g. Cowan and Jonard, 2004; Gulati *et al.*, 2012) or by granting access to multiple paths through which knowledge can reach the firm (König *et al.*, 2011). The same process of knowledge recombination may however set the the premises for the subsequent breaking-up of the network. This is because recombination brings homogeneity into knowledge bases, consequently reducing the incentive for knowledge exchange and thus for alliance formation (Cowan and Jonard, 2004; Gulati *et al.*, 2012). Likewise, in a large network, the number of additional paths to which a firm gets access with an alliance is higher if the alliance is created with a firm which is already part of its component (i.e. if the potential partner is already indirectly connected to the firm). This finding is in perfect agreement with our econometric model, which shows that firms already connected – directly or indirectly – by a path in the network are more likely to form an alliance. In a situation where alliances are costly, this reduces the incentives to maintain bridging ties, thus contributing to the fragmentation of the network into many clusters which are sparsely connected among themselves (see König *et al.*, 2011, for a model generating a similar dynamics). However, a more detailed explanation of the observed rise and fall trend is beyond the scope of the present dissertation. In the next chapters, we will expand the empirical analysis to other collaboration networks in the domain of co-authorship in scientific disciplines. Furthermore, inspired by the findings of our econometric model, we will focus our investigation on the microscopic rules for alliance formation. By developing an agent-based model, we aim at reproducing the emerging topology of the observed collaboration networks and obtaining further insights into the microscopic mechanisms originating such topology, eventually unveiling the complex interdependencies and the mutual feedbacks between network structures and individual firm decisions.

Chapter 3

Similarities among collaboration networks

Summary

Following the empirical findings on R&D networks, in this Chapter we extend our study of trends and patterns on the domain of co-authorship networks in scientific disciplines. Precisely, we select and thoroughly analyze a subset of six representative co-authorship networks obtained from the American Physical Society (APS) databases, spanning from gravitation to interdisciplinary physics (that is, the field including network theory itself). We find that, differently from R&D networks, co-authorship networks do not exhibit any rise-and-fall trend. On the contrary, they exhibit a rise-only trend for most indicators, given the unprecedented expansion that has characterized this domain in the last decades. Some remaining indicators show instead non-constant, fluctuating trends over time, in contrast with most properties of the R&D networks. However, co-authorship networks do show many similarities with R&D networks, both structural and temporal. Our analysis is focused exactly on these universal and robust features. In particular, i. the size of collaboration events (i.e. firms per alliance or authors per paper), ii. the agents' *activity* (i.e. their propensity to engage in a collaboration) and iii. structural communities in the network (beyond the agents' sectoral or geographical positions). Our final goal is to obtain the building blocks for a model capable of reproducing the formation and evolution of different collaboration networks.

Based on M. V. Tomasello, N. Perra, C. J. Tessone, M. Karsai, F. Schweitzer, "The Role of Endogenous and Exogenous Mechanisms in the Formation of R&D Networks", *Scientific Reports*, 4, 5679 (2014). Only the empirical part of the article is included in the present Chapter. M.V.T. contributed to designing the research questions, produced all the statistical analyses and the plots, and gave the major input in writing the manuscript. The analysis of co-authorship networks, not included in the original article, was specifically designed and written by M.V.T. for the present dissertation.

3.1 A brief characterization of co-authorship networks

In this Section we describe the methodology used to build the second family of collaboration networks that we examine in the present dissertation, namely co-authorship networks in scientific disciplines. A co-authorship network is a network whose nodes represent scientific authors, and links represent the papers that they have authored together. The unprecedented, exponential growth characterizing scientific production has also spurred the collection of high volumes of data, often organized in publicly available datasets. A considerable part of the efforts in this kind of research is devoted to the handling of such data and the disambiguation of their entries, as we explain below.

3.1.1 Data and methodology

We use two data sources to construct our co-authorship networks, the American Physical Society (APS) dataset and the Microsoft Academic Search (MSAS) dataset. While the first dataset provides detailed information on papers' abstracts, keywords, received dates, published dates and unique digital identifiers, the second one contains clean and disambiguated information about authors' names and affiliations. Obviously, merging the two sources of information is a necessary procedure to build a co-authorship network.

The datasets The American Physical Society (APS) provides two types of datasets. The first one is a comma-separated value (CSV) table containing all citations within the APS journals, namely, Physical Review Letters, the Reviews of Modern Physics, and all the Physical Review journals, for the period from 1983 to 2010. Each row in the table consists of a pair of Digital Object Identifiers (DOIs) of the citing and the cited papers. A Digital Object Identifier (DOI) is a character string used to identify any electronic document, namely papers published in scientific journals. The DOI is fixed and stable for the entire lifetime of the document, making it a more suitable identifier than the document URL or other kinds of standard identifiers (such as the ISBN). However, as we are not interested in any citation network, we disregard the information about citing and cited papers.

The information we need is contained in the second type of data provided by APS, that is bibliographic meta-data of the papers published in all APS journals. The data is provided in XML format, separately for each journal. Each row contains information about a single paper, including its DOI, journal, title, authors with their affiliations, submission, reviewing, publishing and/or printing dates and PACS codes. The Physics and Astronomy Classification Scheme (PACS) has been developed by the American Institute of Physics specifically for the purpose of classifying scientific papers. PACS codes have a hierarchical structure and are written in the form nn.ab.cd, where nn denotes the research field, ab denotes the subfield and cd denotes the sub-subfield.

The APS data have been processed and the relevant fields stored into a relational database. For the scope of the present study, we keep only the information regarding the paper DOI, printing date, author names, and the first two digits of the PACS numbers, defining the macro research field.

One limitation of the APS dataset is that the authors are identified by strings, often times including inconsistent fields – such as the author's first name initial instead of the full name, missing special characters, or other common spelling mistakes. For this reason, we match the papers in the APS dataset with the MSAS database, where not only the papers, but also the authors are given unique identifiers. Indeed, differently from the APS data, the MSAS data are already fairly well disambiguated with respect to author's first and last names, e-mail address, institution, department and city.

Given that the present study is focused solely on the topological and structural properties of the resulting co-authorship networks, we keep only the authors' unique identifiers, and discard all the information regarding their affiliation and geographical location. Finally, we keep only the entries for which the paper DOI and the remaining available fields are completely matched between the APS and the MSAS datasets. Furthermore, in order to reduce the huge amount of data available and make it compatible with the validation of our models, we select a subset of six relevant PACS numbers, as explained below. By following such procedure, we obtain a total of around 73,000 papers distributed among around 95,000 unique authors.

Construction of the network In order to build a co-authorship network, we assume that every unique author constitutes a node. Then, similarly to the R&D networks (see Chapter 2), we draw a link connecting two nodes every time that a co-authored paper appears in the dataset. From now on, we refer to both R&D alliances or a co-authored papers as collaboration events.

Similarly to R&D alliances, a paper is associated with an *undirected* link, as we do not have any information about the initiator of the scientific collaboration. When a paper is written by more than two authors, all the involved nodes are connected in pairs, thus resulting into a fully connected clique. Following this procedure, the 73,000 papers listed in the dataset result in a total of around 300,000 links in our network representation.

One main difference with the R&D networks is that the authors are not associated with any classification or membership attribute. On the contrary, the classification in co-authorship networks is assigned to the links of the network, i.e. the papers. Indeed, a single paper can be unequivocally assigned to a category (in our case, a PACS number), while a single author can change his/her research subject during his/her career, thus making such a categorization impossible. For this reason, here we build the co-authorship networks in different fields not by selecting the relevant group of nodes, but the relevant group of links, i.e. the papers assigned to a given PACS number.

In order to obtain co-authorship networks that are comparable in size and density with the previously studied sectoral R&D networks, we select the following six representative PACS numbers: 03 (quantum mechanics, field theories and special relativity), 04 (general relativity and gravitation), 42 (optics), 72 (electronic transport in condensed matter), 74 (superconductivity) and 89 (other areas of applied and interdisciplinary physics, that is the field includes network theory itself).

Finally, differently from Chapter 2, we do not consider here any pooled co-authorship network, including all papers in all selected research fields. We follow this approach because of two reasons: i. unlike the pooled R&D network, the overlap between the different research fields is very small, thus giving rise to a network composed of weakly interconnected clusters; ii. the resulting pooled co-authorship network would be computationally difficult to analyze and be used to test the models that we develop in the continuation of the present thesis. We argue that the study of the field specific co-authorship networks already provides statistically significant results and the addition of a pooled network does not improve nor change our results.

3.1.2 Main findings and trends across disciplines

Similarly to Chapter 2, we present here a set of fundamental network indicators, for the computation of which we assume that every link is terminated 3 years after its formation. Indeed, a collaboration established through a scientific paper is intrinsically impossible to have a predetermined duration, thus forcing us to make such an assumption. However, similarly to R&D networks, our results are robust to the length of such duration. In any case – as we explain below – we will shift the focus to other relevant network quantities, that are more robust and stable over time and across domains, thus allowing a meaningful modeling and understanding of different collaboration networks.

Network size and density. We present in Fig. 3.1 a visual representation of the six co-authorship networks that we have selected, plotting all network snapshots in the years 1989, 1993, 1997, 2001, 2005 and 2009. All networks are displayed using the Fruchterman-Reingold algorithm (see Chapter 2 for more details).

The plots suggest that all research fields have experienced a network growth, without displaying the rise-and-fall trend typical of R&D networks. The growth in network size is always associated with the decrease in network density, similarly to R&D networks, as we show in Fig. 3.2, meaning that the addition of new nodes is the driving force for the network growth.

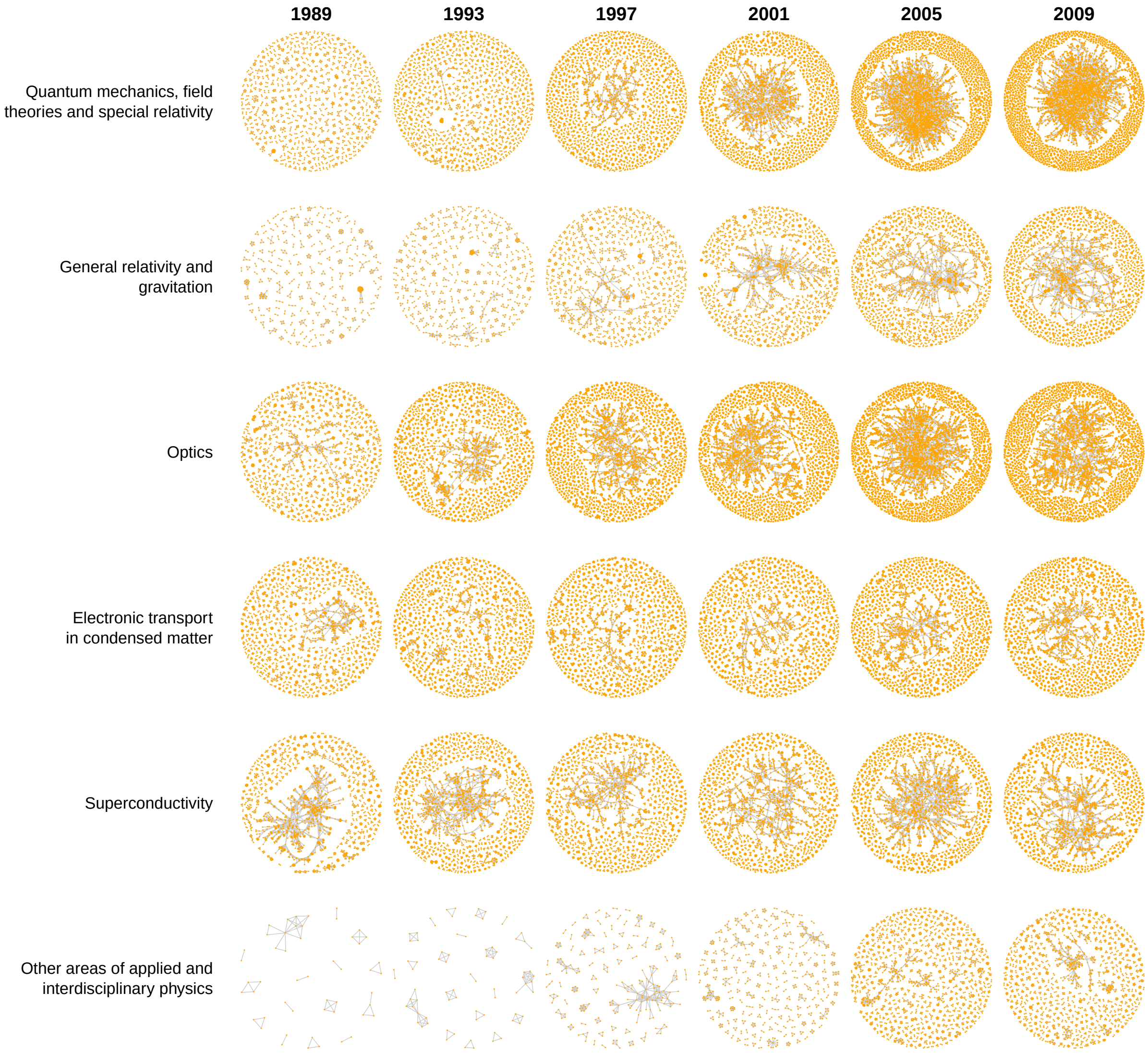

1989
1993
1997
2001
2005
2009
Quantum mechanics, field theories and special relativity
General relativity and gravitation
Optics
Electronic transport in condensed matter
Superconductivity
Other areas of applied and interdisciplinary physics

Figure 3.1: Network snapshots in 1989, 1993, 1997, 2001, 2005 and 2009 for the six representative co-authorship networks that we have selected for the present study. All network layouts are computed using the Fruchterman-Reingold algorithm.

We find that all networks display a monotonous increasing trend for their size, with small field-related differences – for instance, the superconductivity field experiences a sort of saturation after 1995, while the applied and interdisciplinary physics field experiences an exponential growth after the year 2000. All fields seem to exhibit a small decrease in network size in the year 2009. Differently from R&D networks, where the rise-and-fall trend has been proven to be real and consistently recorded by the data (Schilling, 2009), here we argue that such final decrease is simply due to incomplete data towards the end of the observation period. The increasing trend in all fields of scientific production is a well known and documented phenomenon, and the datasets employed here were probably not consolidated yet at the moment of their usage.

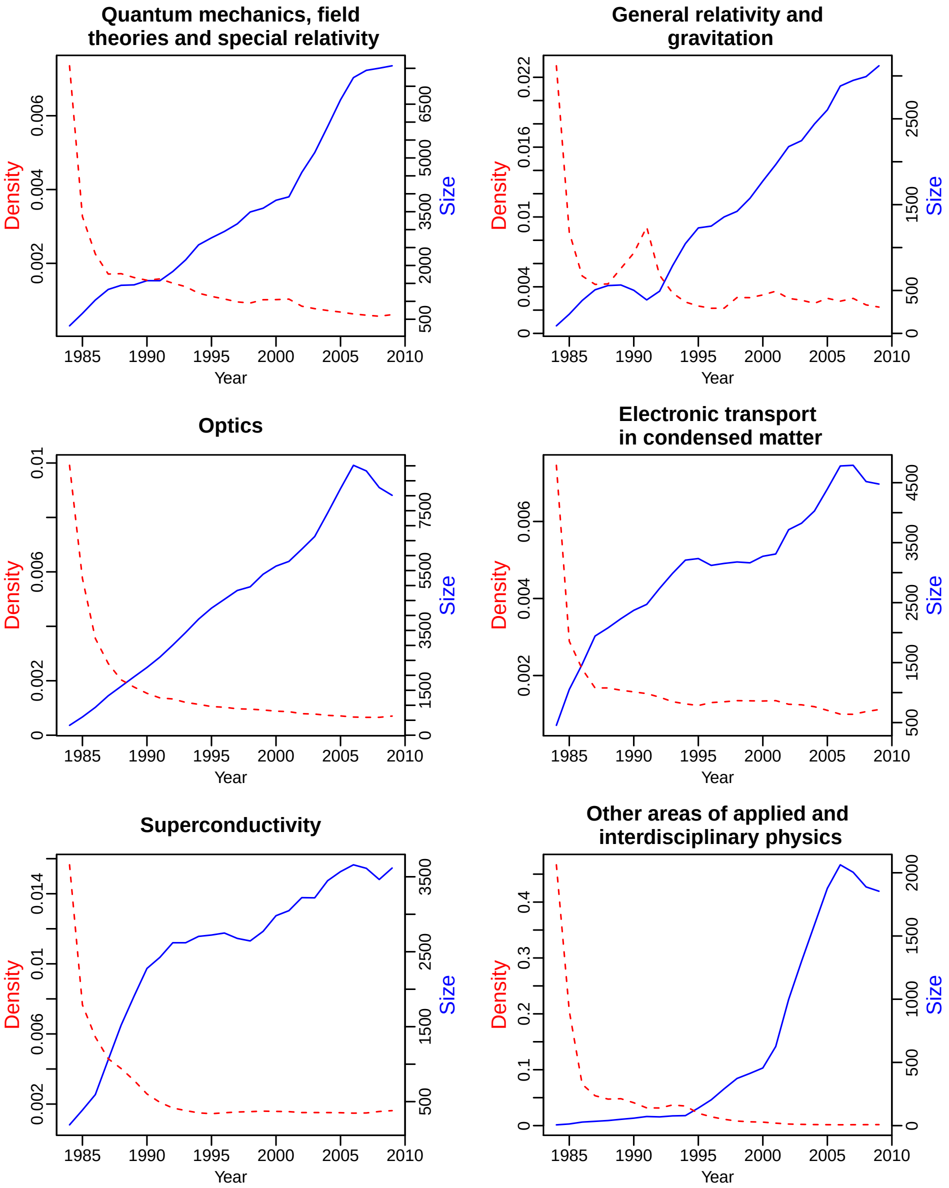


Figure 3.2: Time-evolution of size (solid blue line, right axis) and density (dashed red line, left axis) for our six representative co-authorship networks.

Giant component and degree heterogeneity. Similarly to the R&D networks, the growth of co-authorship networks is associated with the emergence of a giant component, as we report in Fig. 3.3. The only notable exception is represented by the field of applied and interdisciplinary physics, where the increase in size has resulted in a bigger fragmentation of the network and a decrease in the size of the main network component. In all other fields, although showing some fluctuation over time, the giant components have sizes ranging from 30% to 55% of the entire network.

However, differently from R&D networks, such growth is not associated with an increasing inequality of the node degrees in the network, as we show in Fig. 3.3. The quantity we

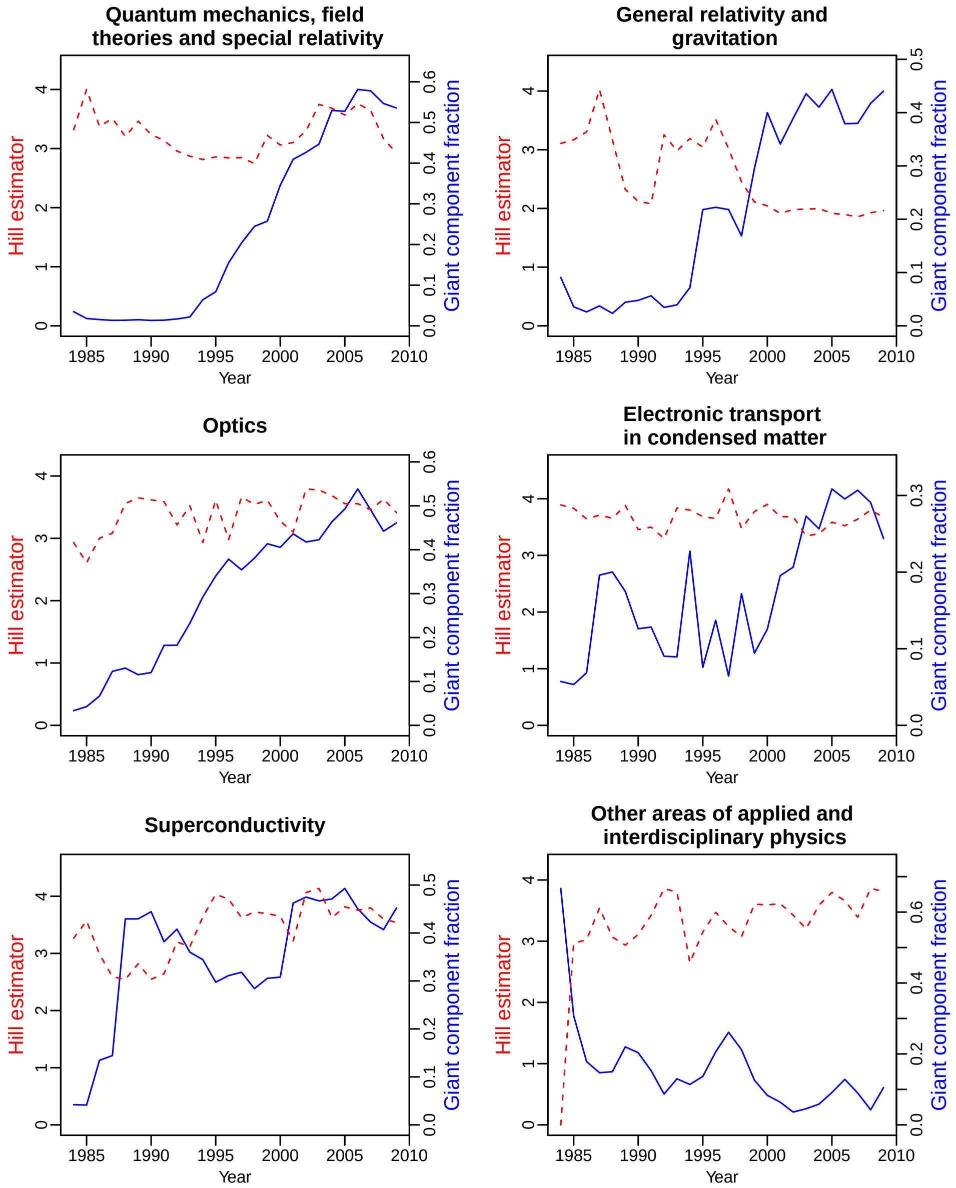


Figure 3.3: Time-evolution of the giant component fraction (solid blue line, right axis) and the degree distribution's Hill Estimator (dashed red line, left axis) for our six representative co-authorship networks.

plot is the Hill Estimator (HE) of the different degree distributions. The smaller its value, the more heterogeneous and right-skewed the corresponding degree distribution is (see Chapter 2 for more details). For a comparison, a network generated by a simple preferential attachment mechanism would display a degree distribution with HE equal to 3. We find that all research fields have degree distributions with stable HE after 1985, around values compatible with the preferential attachment mechanism.

Remarkably, the co-authorship network in the general relativity and gravitation field exhibits a HE that stabilizes around 2, signaling a heavy right tail in its degree distribution.

This fact can be explained by the larger and larger number of authors that is typically involved in the writing of scientific papers in this field. All other research fields show small fluctuations over time in their degree distributions' HE. In many cases, these fluctuations bring the values of HE slightly above 3, meaning that the corresponding degree distributions are narrower than the one generated by a preferential attachment mechanism – differently from most of the sectoral R&D networks.

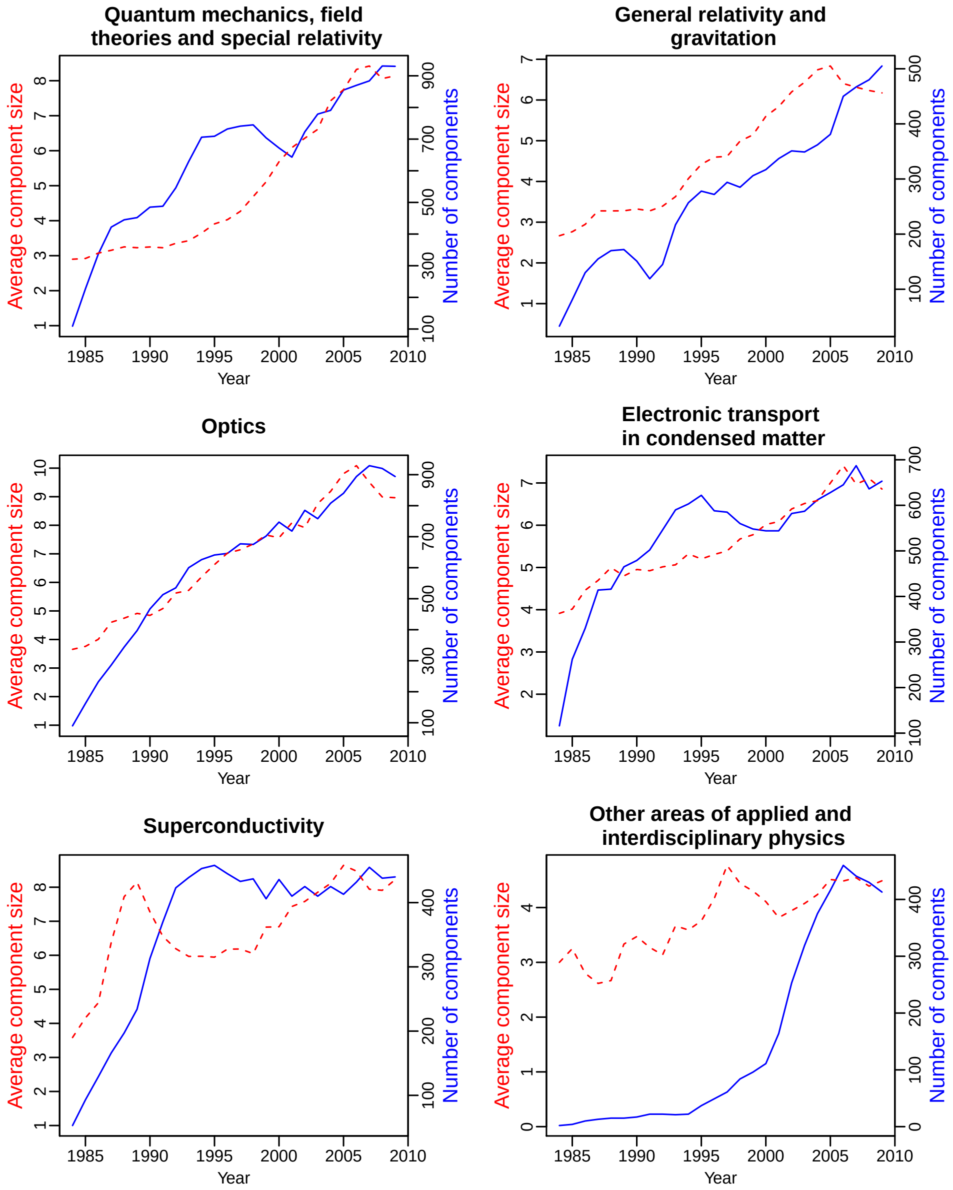


Figure 3.4: Time-evolution of the number of network components (solid blue line, right axis) and the average component size (dashed red line, left axis) for our six representative co-authorship networks.

Network components. In order to have a more detailed picture of the network structure, we study the evolution of the number of disconnected components and the average component size in every collaboration network. Our results are reported in Fig. 3.4. We find that the number of network components scales with the network size, meaning that the growth we have observed is associated with the addition of more and more disconnected components. Again, the co-authorship network in superconductivity exhibits a sort of saturation effect after 1995, and the co-authorship network in applied and interdisciplinary physics exhibits an exponential growth after 2000.

However, the average components size increases over time as well for every co-authorship network, meaning that such networks are overall more and more connected – consistently with the emergence of giant components. Yet again, the network in the superconductivity field exhibits a sort of saturation after 1990, combined with a negative fluctuation around year 1995, and the network in applied and interdisciplinary physics exhibits only a small increase in the average component size – consistently with the lack of a giant component in this research field.

Degree assortativity and small world properties. Finally, we study the evolution of two characteristic network indicators, the assortativity mixing coefficient – measuring the degree-degree correlations in the network – and the small world quotient – measuring the extent to which the network has a lower average path length and a higher clustering coefficient than a corresponding randomly generated network. For more details on these coefficients, see Chapter 2. Our results for both indicators are reported in Fig. 3.5.

We find that the assortativity mixing coefficient is always positive for every collaboration network, which then exhibit the typical features of social networks (Newman, 2003). This means that, in co-authorship networks, nodes with small degrees tend to be connected with other small-degree nodes, while nodes with high degrees tend to be connected with other high-degree nodes. In particular, we find that the assortativity mixing coefficient is always greater than 0.2 for all co-authorship networks in all time periods, sometimes reaching peaks as high as 1; however, such degree-degree correlations do not display any regular trend as a function of time. Such finding constitutes a remarkable difference with the sectoral R&D networks, which instead exhibit the negative degree-degree correlations typical of technological networks (see Chapter 2).

The small world quotient does not exhibit any particular trend as a function of time. We find that all curves show considerable fluctuations during all time periods; however, towards the end of our observation period, they stabilize around values between 200 and 800. This means that – similarly to most of the R&D networks (see Chapter 2) – all co-authorship networks exhibit significant small world properties, i.e. overall low average path lengths and high clustering coefficients.

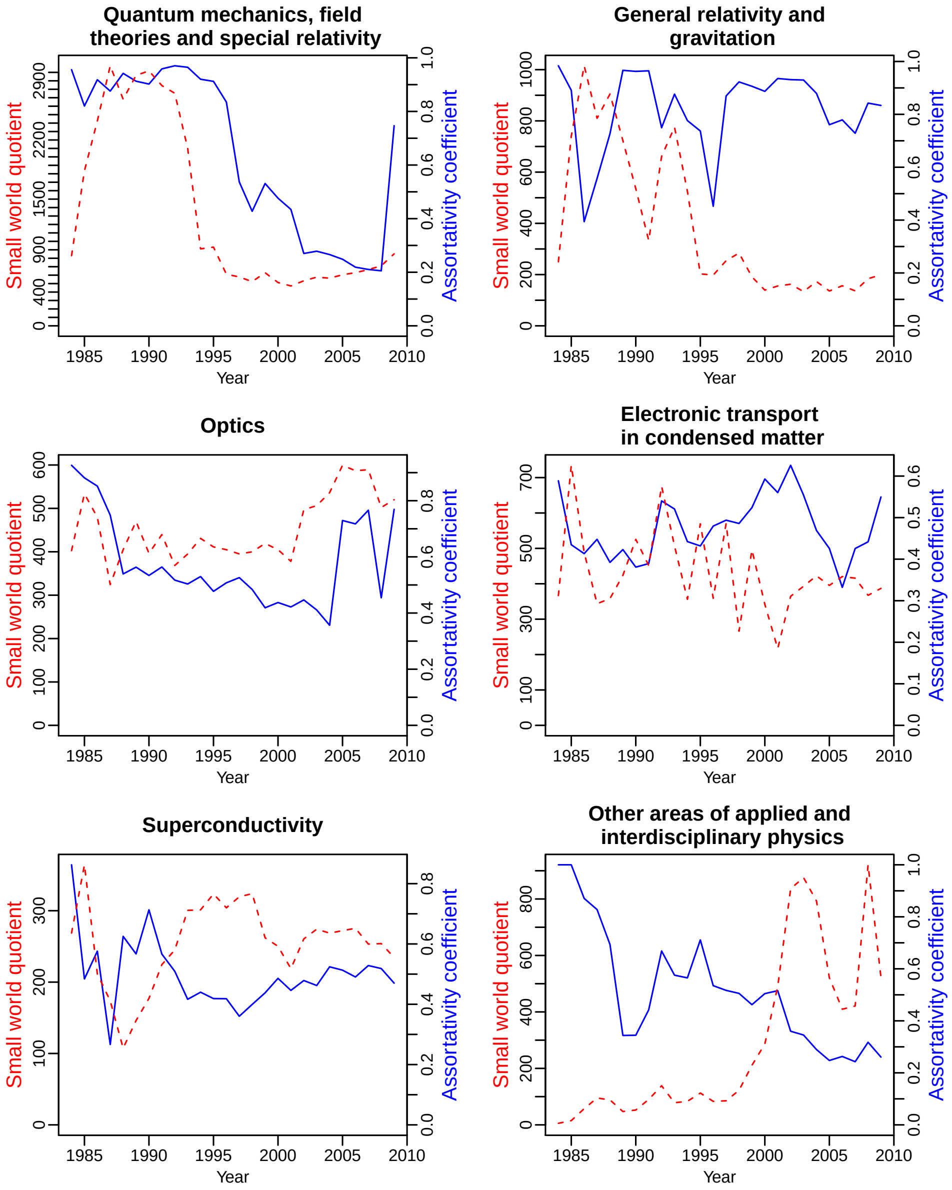


Figure 3.5: Time-evolution of the mixing assortativity coefficient (solid blue line, right axis) and the small-world quotient (dashed red line, left axis) for our six representative co-authorship networks.

3.2 Robust network features across domains

The analysis of R&D networks and co-authorship networks has shown us some controversial findings. On the one hand, R&D networks display rise-and-fall trends for most indicators, displaying a “golden age”, mainly characterized by larger size, smaller density, more and larger network components, an increased degree heterogeneity across nodes, and small world architectures. The only indicator showing heavy fluctuations is the assortativity coefficient (i.e. degree-degree correlations) in the networks.

On the other hand, co-authorship networks are characterized by generally rising trends in terms of size, component number and average component size, associated with decreasing trends for the network density, and non-constant, remarkably fluctuating trends for degree heterogeneity across nodes, assortativity and small world properties.

Such differences prevent us from arguing that every time a collaboration network evolves – both in the domain of R&D alliances and writing of scientific papers – such growth results in the same macroscopic structures. Instead, being those structures the result of some specific microscopic mechanisms of link formation, we can conclude that such mechanisms must differ from network to network.

Building blocks for an agent-based model. One of the aims of the present dissertation is to develop an agent-based model including the minimum possible number of microscopic rules, that is able to reproduce the topology of real collaboration networks. This means that we still need to search for a minimal set of features, patterns or rules that are robust across domains, and that can be therefore used as building blocks of our agent-based model. Our aim is then to reproduce all the observed sector-related or field-related differences in real collaboration networks, by tuning some of the parameters of such model.

The network features that we have studied and reported so far, rather than being a starting point, will be instead used to validate such model and fine-tune its parameters in the different collaboration networks, in both the R&D and the co-authorship domains. Such procedure will be explained in detail in Chapter 4.

In the remainder of the present Chapter, we study a different set of features on real collaboration networks. Such features are more elementary and primitive than the ones previously studied, thus representing more suitable basic blocks for our future agent-based model.

3.2.1 Size of collaboration events

The first basic, universal feature under our examination is the size of collaboration events in collaboration networks. With respect to the pooled R&D network, we find that the SDC alliance dataset exhibits a right-skewed distribution of number of partners per alliance event. Most of the collaborations (93%) are stipulated between two partners, but some alliances – the so-called *consortia* – involve three or more partners. The distribution of the number of firms per alliance event, that we show in Fig. 3.6, spans one order of magnitude. In Fig. 3.7 we report such distribution for the six largest industrial sectors, showing that this feature holds for all sectoral R&D networks, with only small differences in the tails of the respective distributions.

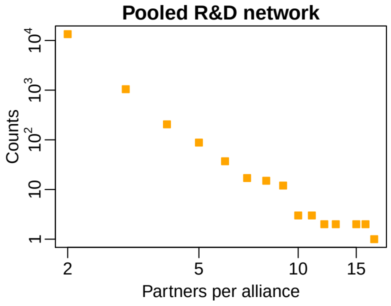


Figure 3.6: Distribution of the number of partners per alliance, as measured from the SDC alliance dataset.

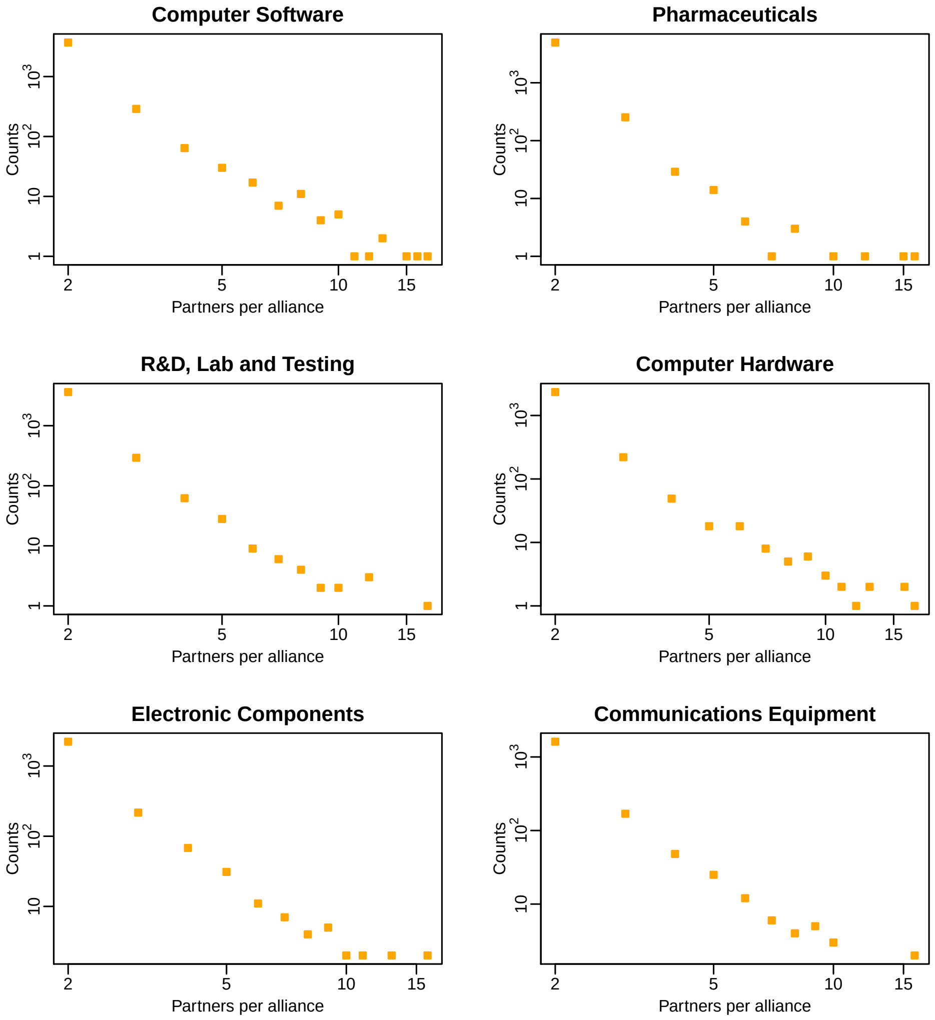
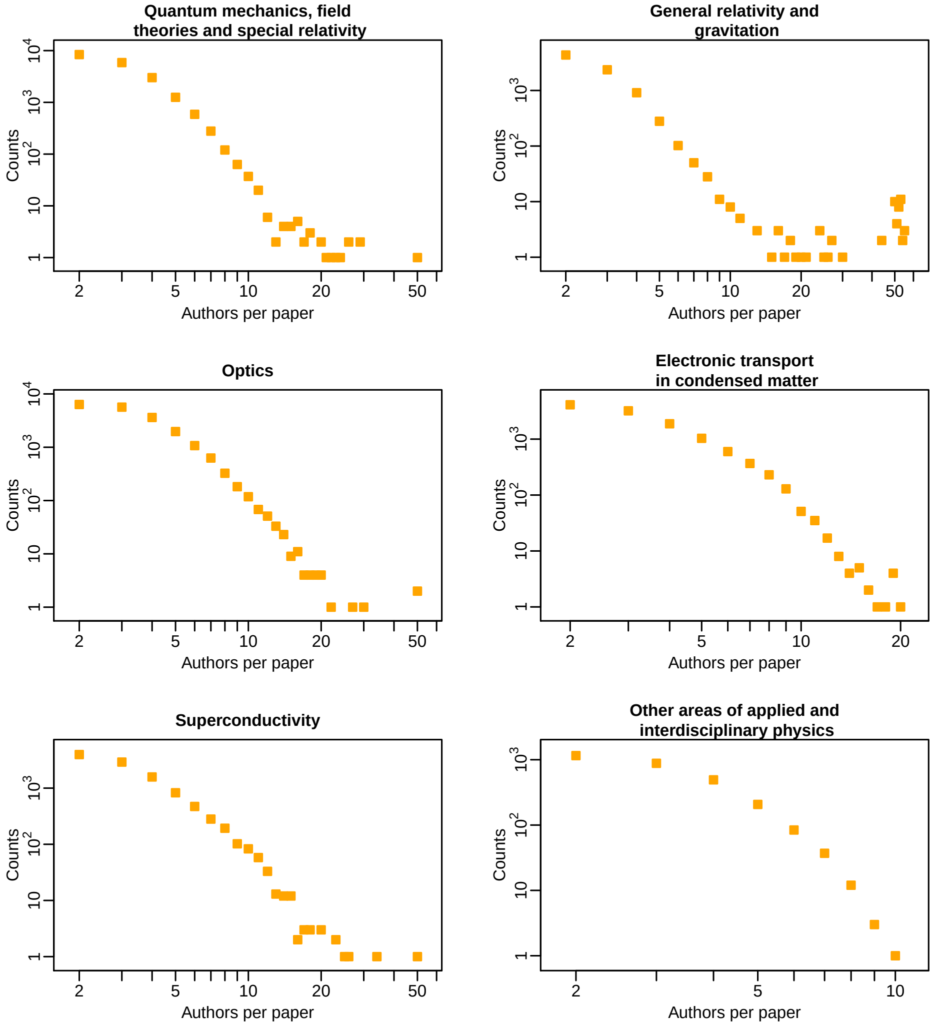


Figure 3.7: Distribution of the number of partners per alliance for the six largest industrial sectors, as measured from the SDC dataset.

The typical right-skewed distribution of agents per collaboration event holds also for all the co-authorship networks that we analyze. Our results are reported in Fig. 3.8. It should be noted that, obviously, many papers are written by only one author; however, we exclude such events from our collaboration network representation – including such authors would generate isolate nodes. Therefore, the counts start from 2 in all of our plots.



Quantum mechanics, field theories and special relativity

General relativity and gravitation

Optics

Electronic transport in condensed matter

Superconductivity

Other areas of applied and interdisciplinary physics

Figure 3.8: Distribution of the number of authors per paper for our six representative co-authorship networks, as measured from the APS-MSAS datasets.

Differently from the R&D networks, the co-authorship networks exhibit a larger degree of variability among fields. First of all, the typical number of authors per paper strongly depends on the field. To give an example, the field of applied and interdisciplinary physics is characterized by significantly fewer authors per paper (at most 10) than the field of general relativity and gravitation (whose right tail reaches 55 authors per paper). In particular, we find that the distribution in this last field is characterized by a bimodal behavior, with a bump in counts around 50 – this is due to the publishing activity in the gravitation subfield, which typically requires such large number of authors to conduct a single experiment. However, also the other research fields (with the only exception of applied and interdisciplinary physics) exhibit distributions with heavy right tails.

In addition, such distributions seem to have an exponential or stretched exponential form, rather than the typical power law trend of the R&D networks. However, our scope is not to study the origin of these distributions, nor determine their functional form, nor reproduce them. They rather constitute a starting point of our future agent-based model, that will instead reproduce other and more sophisticated network features. Therefore, we just record these distributions in order to feed such empirical inputs to the model.

3.2.2 Agents' activity

Another distinctive measure we introduce and analyze in this study is the agents' *activity* distribution (Perra *et al.*, 2012). Developed in the field of temporal networks (Holme and Saramäki, 2012), the activity has already been studied on various datasets, such as online microblogging, actor/movie networks and co-authorship networks as well. However, to the best of our knowledge, no previous work has measured this quantity on a set of real firms involved in R&D alliances by using empirical data. This certainly represents a logical consequence of such recent developments in temporal networks.

We define the empirical *activity* $a_{i,t}^{\Delta t}$ of an agent i at time t, over a time window Δt, as the number of collaboration events $e_{i,t}^{\Delta t}$ involving the agent i in the time window Δt ending at time t, divided by the total number of collaboration events $E_t^{\Delta t}$ involving *any* agent in the network during the same time period:

$$a_{i,t}^{\Delta t} = \frac{e_{i,t}^{\Delta t}}{E_t^{\Delta t}}. \tag{3.1}$$

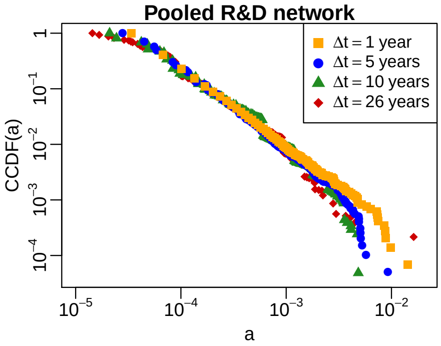


Figure 3.9: Complementary cumulative distribution function (CCDF) of the empirical firm activities in the pooled R&D network, measured on the SDC dataset with 4 different time windows Δt of 1, 5, 10 and 26 years. When the time window is shorter than 26 years (the entire dataset observation period), we compute the activity by shifting the time window in 1-year increments and then we average the results.

The activity expresses the probability that an agent takes part in an arbitrary collaboration event occurring in a given time window. We test four time window lengths Δt equal to 1, 5, 10 and 26 years for both the SDC alliance dataset and the APS-MSAS co-authorship dataset. We find that all agent activity distributions are virtually independent of the chosen Δt. We report our findings for the pooled R&D network in Fig. 3.9. In this and the totality of next plots, we make use of the complementary cumulative distribution function (CCDF), similarly to the approach adopted in Chapter 2, because it is more stable and gives better visual representations compared to the simple probability density function.

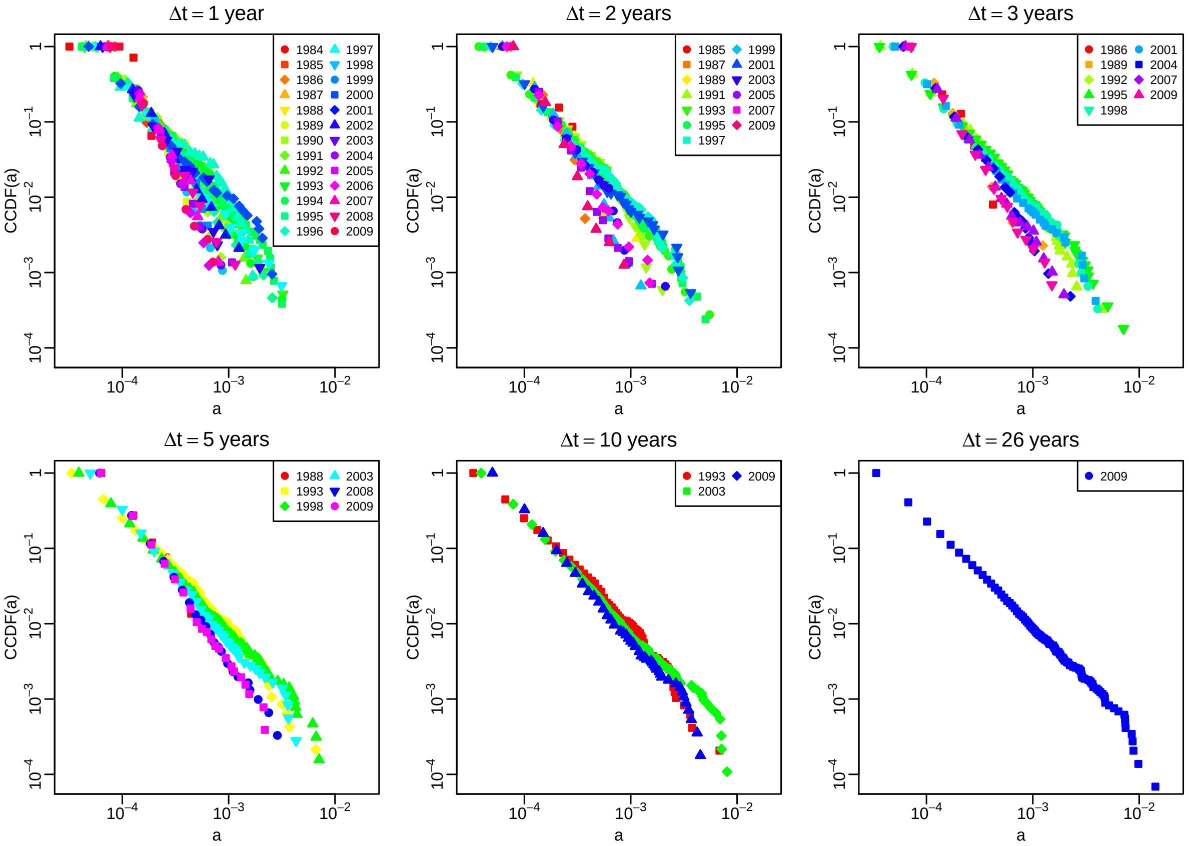


Figure 3.10: Complementary cumulative distribution function (CCDF) of the empirical firm activities in the pooled R&D network, measured on the SDC dataset with 6 different time windows Δt of 1, 2, 3, 5, 10 and 26 years. When the time window is shorter than 26 years, we shift such time window along the observation period and show the corresponding activity CCDF.

We find that the firm activity distributions are right skewed and dispersed over several orders of magnitude, as in many other social and technological systems (Barabasi, 2005; Barabasi and Albert, 1999; Pastor-Satorras *et al.*, 2001). Contrary to most of the R&D network indicators, that display strong variability and dependence on time (see Chapter 2), the activity is a stable attribute that can be assigned to firms and effectively estimate their propensity to engage in new alliances. Indeed, empirical firm activities are robust also with respect to the time t at which they are measured: shifting the time window – of any length Δt – along the 26 years reported in the dataset does not affect the results, as we show in Fig. 3.10 for the pooled R&D network.

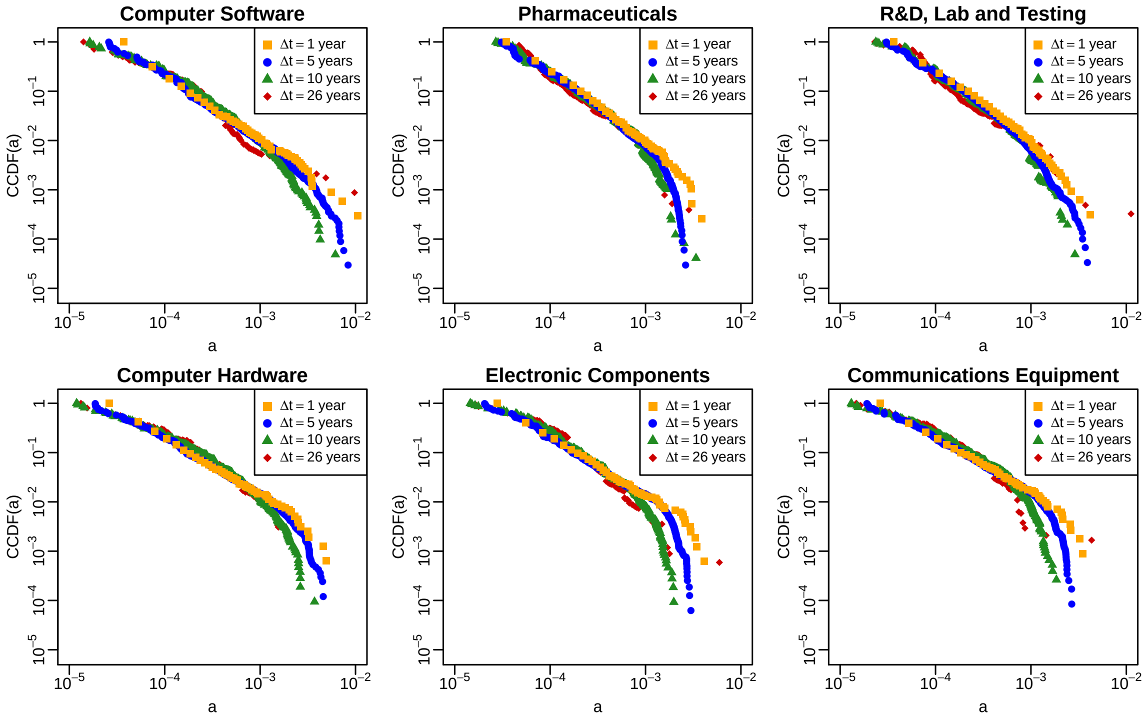


Figure 3.11: Complementary cumulative distribution function (CCDF) of the empirical firm activities, measured for the six largest industrial sectors in the SDC dataset.

In addition, we find that the activity distribution is robust to the sectoral classification of the firms. In Fig. 3.11 we show the empirical firm activity distributions (computed on four different time windows) for the nine largest sectoral R&D networks.

The activity distributions for the authors in our six representative co-authorship networks are reported in Fig. 3.12. We find that the trend of all distributions is robust with respect to both the time window and the research field, similarly to R&D networks. We use the same time window lengths and the same observation period as the R&D networks, to allow for a straightforward comparison.

In order to prove the robustness of the empirical activity in co-authorship networks to the time t at which it is measured, we report in Fig. 3.13 the effects of shifting the time window – of any length Δt – along the 26 years of observation. For the sake of clarity and brevity, we analyze here only one representative co-authorship network – the one in applied and interdisciplinary physics – and four different time windows – of length 1, 5, 10 and 26 years. Similarly to R&D networks, and to the other co-authorship networks, shifting the time window does not affect the shape of the activity distributions.

To sum up, we find that agents in collaboration networks are endowed with an activity, an attribute measuring their propensity to engage in a collaboration. The distributions

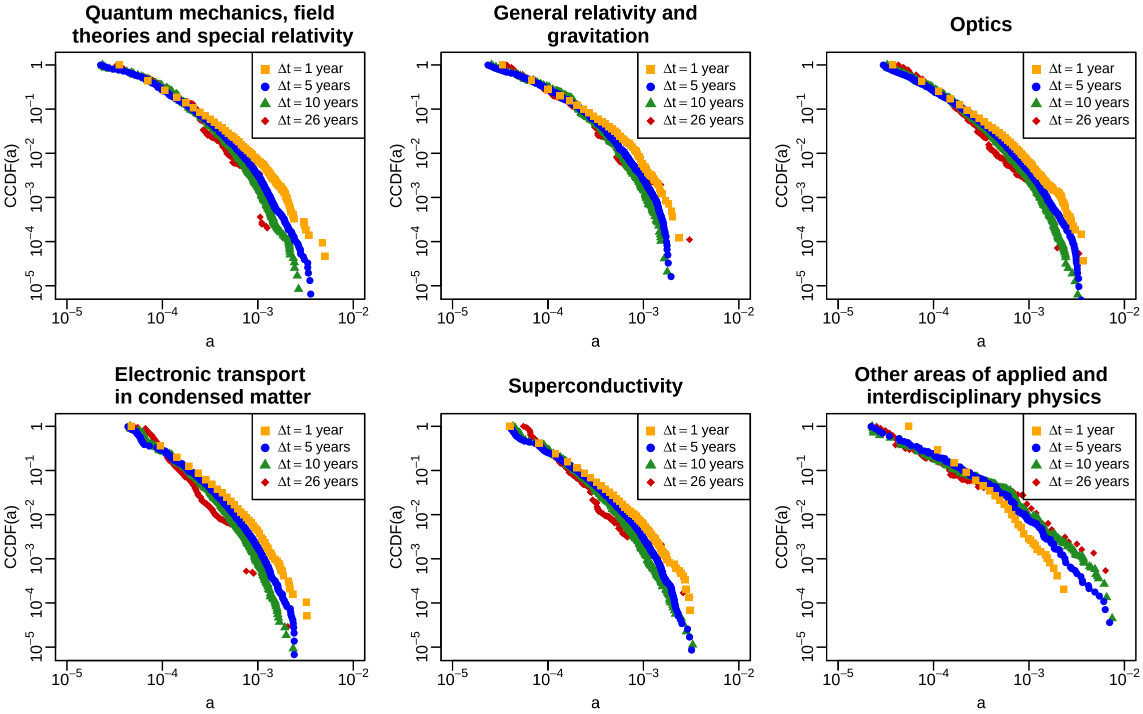


Figure 3.12: Complementary cumulative distribution function (CCDF) of the empirical author activities, measured for the six selected co-authorship networks in the APS-MSAS datasets.

of such agent activities in R&D and co-authorship networks is right skewed and dispersed over several orders of magnitude, as in many other social and technological systems.

We also find that the agent activities are very stable across domains and over time, thus making them perfect candidates for stable agent attributes, representing their propensity to engage in a collaboration event. Similarly to the distribution of number of agents per collaboration, we record also these distributions as empirical inputs for our future agent-based model.

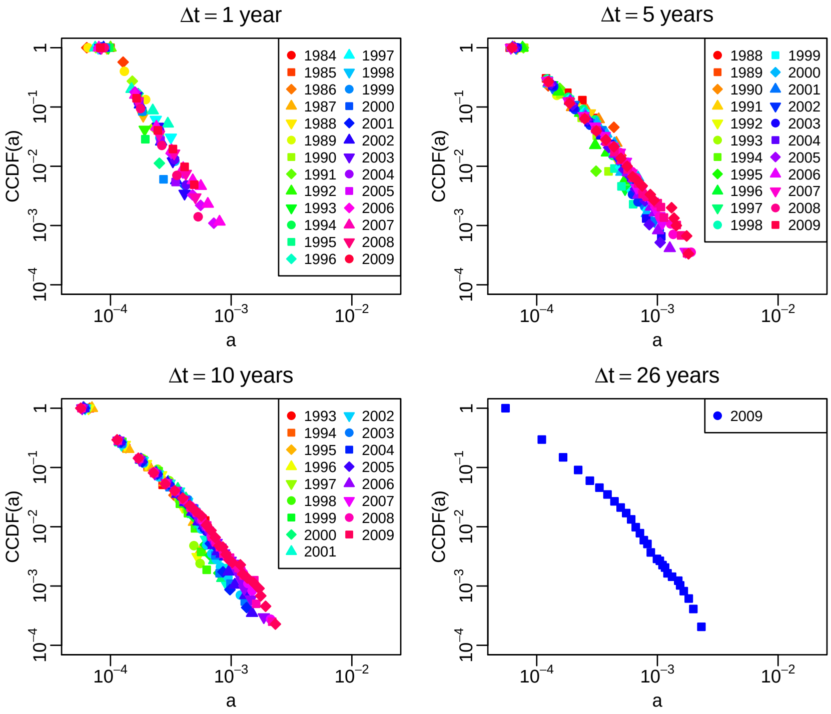


Figure 3.13: Complementary cumulative distribution function (CCDF) of the empirical author activities in the applied and interdisciplinary physics co-authorship network, measured on the APS-MSAS datasets with 4 different time windows Δt of 1, 5, 10 and 26 years. When the time window is shorter than 26 years, we shift such time window along the observation period and show the corresponding activity CCDF.

3.2.3 Communities and labels

Finally, we turn our attention to the modular properties of collaboration networks. It has been acknowledged that networks, in many different domains, are organized in modules or clusters, characterized by groups of tightly connected nodes (Fortunato, 2010; Newman and Girvan, 2004). We find that both R&D and co-authorship networks are not an exception (see Chapter 3).

In R&D networks, interestingly, the formation of such clusters is not totally explained by factors like the firms' industrial sectors or their geographical distribution (Rosenkopf and Padula, 2008). Our previous analyses (see Chapter 2) have also shown that the link formation is explained as well by the belonging to the same country and sector as previous network structures. Indeed, firms belonging to different sectors and located in different countries can populate the same network cluster. However, clusters in R&D networks have never been theoretically defined; they have been only empirically detected by means of simple K-means algorithms and used to obtain rough indications about the inter-firm alliance activity (Rosenkopf and Padula, 2008).

Here, we perform a community detection on the pooled R&D network by employing a widely used algorithm (Infomap) and report our findings in Fig. 3.14. The Infomap algorithm detects structural clusters based on the probability flow of random walks in the network (Rosvall and Bergstrom, 2008). We detect the presence of approximately 3,500 clusters in the network, whose size distribution appears to be dispersed and right skewed, displaying a maximum cluster size of about 200 firms and a minimum cluster size of 2.

In Fig. 3.14 we also provide a representation of the pooled R&D network; for the sake of visualization, we consider only the 30 largest firm clusters and depict them by grouping the corresponding nodes in 30 distinct regions of the plot area. It should be noted that such visual representation strongly differs from the one that we have provided in Fig. 2.1. First, here we do not depict sectors with different colors, because we are interested only in the structural network clusters. Second, the pooled R&D network we represent here is *cumulative*, i.e. it includes all alliances reported in the dataset, without assuming their termination after three years. This choice is consistent with the microscopic rules that we will set for our agent based model. Even though such model will be aimed at reproducing cumulative structures and patterns, other dynamical and temporal microscopic indicators will be tested, thus not affecting the validity of our predictions. More explanations and details will follow in Chapter 4.

Finally, we compute the modularity score Q of the pooled R&D network, to quantify the goodness of such division of the network in clusters. Such coefficient (Newman, 2010) is defined such that $Q = 1$ in case of a perfectly modular network, where links are formed only within the same cluster. Likewise, $Q = -1$ for a perfectly anti-modular network, where links connect only nodes belonging to distinct clusters, and $Q = 0$ for a network

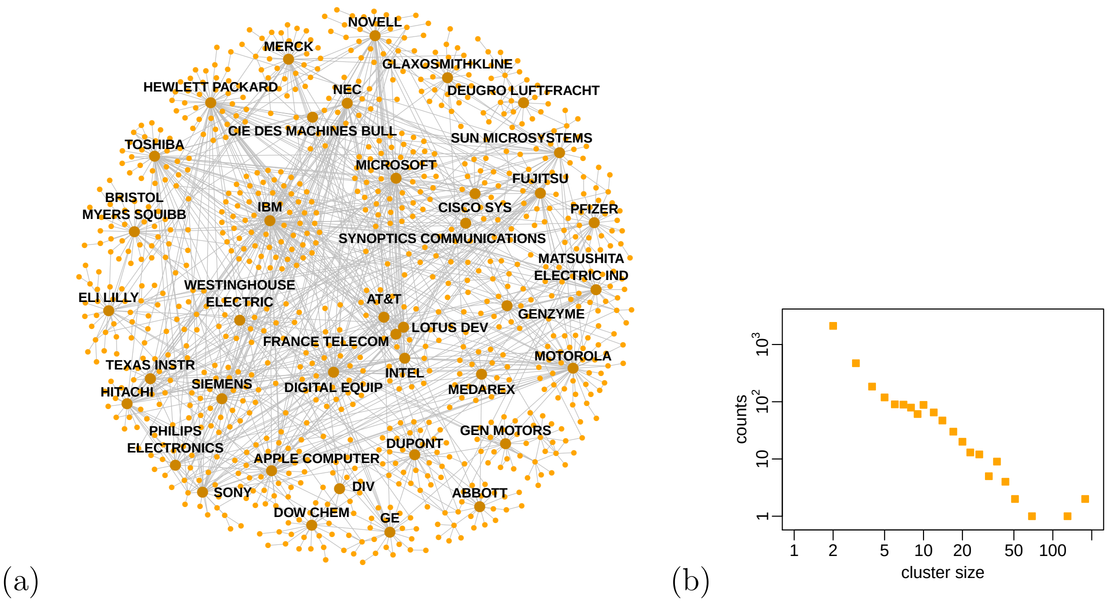


Figure 3.14: (a) Visual representation of the empirical R&D network (we use the Fruchterman-Reingold layout (Fruchterman and Reingold, 1991)), considering only the 30 largest clusters detected by the Infomap algorithm. Distinct clusters are represented by grouping nodes in distinct regions of the plot area. The highest degree nodes are also highlighted: some big companies in several industrial sectors, together with their respective clusters, are clearly visible in the plot. (b) Size distribution of the network clusters.

where links are formed at random. For more details on the Q coefficient, see Chapter 2. For the pooled R&D network, we observe a value of 0.679, remarkably high not only if compared to other examples of real networks (Newman, 2004a), but also if compared to the Q values that we have obtained in Chapter 2, where an industrial sectoral division has been used. This means that the belonging of firms to different sectors does not reproduce the topological network clusters that we detect.

Next, we apply the same exact procedure to a set of representative sectoral R&D networks, as well as co-authorship networks. We report all basic statistics in Table 3.1. However, for the sake of brevity, we report the visual network representations and the cluster size distributions only for two representative examples, the Pharmaceuticals sectoral R&D network (Fig. 3.15) and the co-authorship network in applied and interdisciplinary physics (Fig. 3.16). The results for the remaining collaboration networks are reported in Appendix B.

We find that all collaboration networks, both in the R&D and in the co-authorship domains, are characterized by high modularity scores. Precisely, all the Q scores originated by an Infomap community detection are significantly higher than the equivalent scores on randomly generated networks with the same degree sequence, especially in the domain of co-authorship networks. By following such approach (Reichardt and Bornholdt, 2006), we

	N	E	*Links*	*Clusters*	Q	Q^{rand}
Pooled R&D network	14,561	14,829	21,572	3,561	0.679	0.570 ± 0.001
Sectoral R&D networks						
Pharmaceuticals (SIC 283)	3,829	5,277	6,019	860	0.607	0.438 ± 0.002
Computer hardware (SIC 357)	1,582	2,672	4,047	783	0.623	0.502 ± 0.002
Communications equipment (SIC 366)	1,133	1,888	2,726	749	0.653	0.461 ± 0.002
Electronic components (SIC 367)	1,615	2,574	3,756	302	0.502	0.311 ± 0.002
Computer software (SIC 737)	3,381	4,134	5,862	354	0.531	0.333 ± 0.002
R&D, laboratory and testing (SIC 873)	3,188	4,032	5,364	256	0.527	0.317 ± 0.003
Co-authorship networks						
Quant. mech., field theories, spec. relativity (PACS 03)	21,501	19,647	56,111	3,029	0.779	0.2344 ± 0.0004
General relativity and gravitation (PACS 04)	8,294	8,158	32,513	1,207	0.795	0.128 ± 0.016
Optics (PACS 42)	27,436	20,105	94,961	2,853	0.794	0.195 ± 0.002
Electronic transport in condensed matter (PACS 72)	19,492	11,687	55,818	2,411	0.832	0.2609 ± 0.0004
Superconductivity (PACS 74)	14,920	10,541	52,615	1,663	0.769	0.208 ± 0.003
Other applied and interdisciplin. physics (PACS 89)	4,881	2,873	8,777	966	0.920	0.395 ± 0.001

Table 3.1: Modular properties for the pooled R&D network, the six largest sectoral R&D networks, and the six representative co-authorship networks. For all domains, we consider the respective cumulative networks, i.e. the networks obtained by keeping all the links at any time. For each network, we report the number of nodes N, of collaboration events E, of resulting links in our network representation, of clusters detected by the Infomap algorithm, the modularity score Q of the network, and (as robustness check) the modularity score Q^{rand} obtained in a set of 100 randomly generated networks with the same size and degree sequence as the network under examination.

can safely conclude that – in every collaboration network that we test – such high Q values are indicative of a real modular structure, and not a simple artifact of the network's size and density.

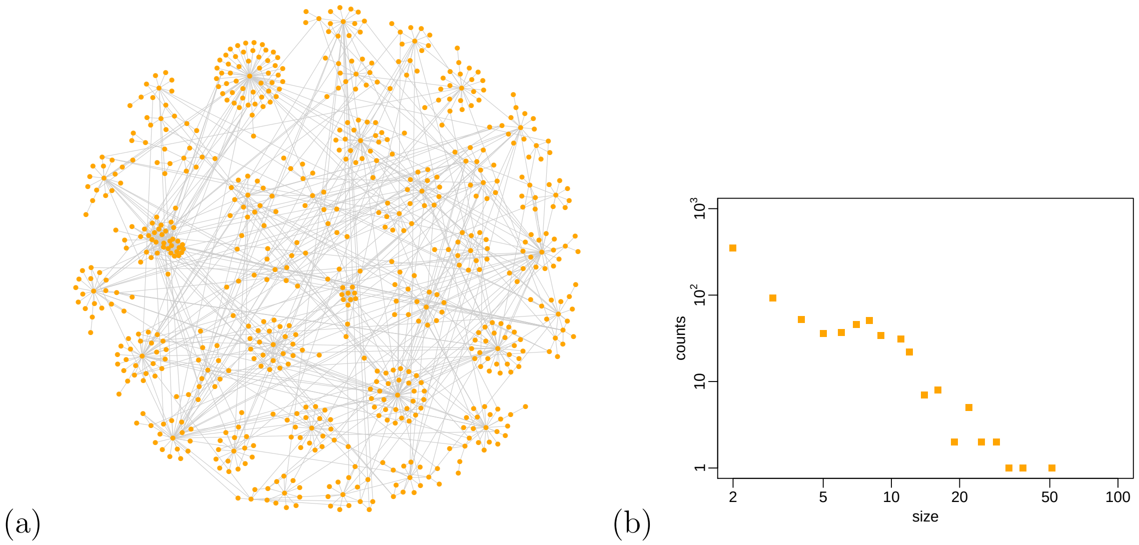


Figure 3.15: (a) Visual representation of the Pharmaceuticals sectoral R&D network, considering only the 30 largest clusters detected by the Infomap algorithm. Distinct clusters are represented by grouping nodes in distinct regions of the plot area. (b) Size distribution of the network clusters.

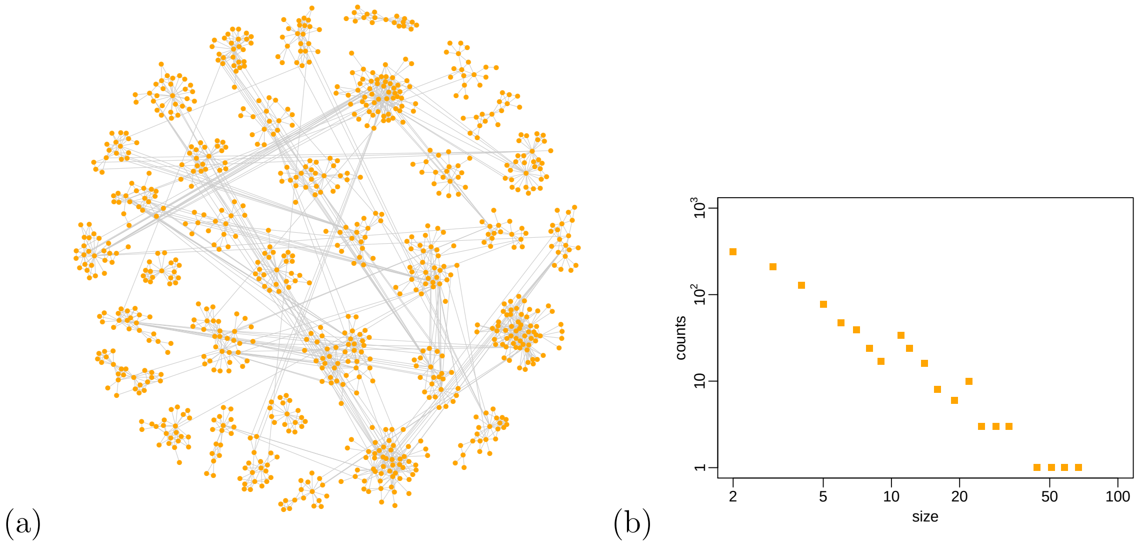


Figure 3.16: (a) Visual representation of the co-authorship network in applied and interdisciplinary physics, considering only the 30 largest clusters detected by the Infomap algorithm. Distinct clusters are represented by grouping nodes in distinct regions of the plot area. (b) Size distribution of the network clusters.

In conclusion, by performing a community detection algorithm, we have found that collaboration networks are characterized by modular structures. This finding is robust across domains, and supported by the evidence of significantly lower modularity scores on randomly generated networks with preserved degree sequence – i.e. the modularity is not an artifact of the specific networks' degree sequences. However, differently from the distributions of agent activity and number of agents per collaboration, we will not use such result as a building block for our agent based model. The network modularity is a complex, emerging phenomenon of the evolution of collaboration networks, and we will rather use it as a criterion to validate the predictions of our model.

We argue instead that such modular structures are indicative of some specific microscopic rules of strategic link formation, probably involving the presence of an agent membership attribute. This intrinsic membership attribute goes beyond the sectoral, geographical or research field belonging of the agents, and causes them to form dense network clusters, albeit allowing a certain level of inter-cluster connections. This hypothesis is in line with the findings reported by Yang and Leskovec (2012), that have tried to identify the presence of communities based on ground truth in real networks. Therefore, we use such concept of membership attribute as a microscopic rule for our agent-based model, leaving the modularity and the emergence of clusters as a validity test.

3.3 Discussion

The implications of the analyses carried out in the present Chapter are twofold. First, we have extended the investigation of network trends and patterns from R&D to co-authorship networks in scientific disciplines. We find that co-authorship networks are characterized by similar network structures to the R&D networks, i.e. the emergence of giant connected components, heterogeneous degree distributions, small world properties, and a positive degree assortativity coefficient (this is in agreement with the pooled R&D network, but in contrast to the sectoral R&D networks, which are generally characterized by negative degree-degree correlations).

However, differently from R&D networks, co-authorship networks do not exhibit any rise-and-fall trend. On the contrary, they are characterized by generally rising trends over the last three decades, in terms of size, component number and average component size, associated with decreasing trends for the network density, and non-constant, remarkably fluctuating trends for degree heterogeneity across nodes, assortativity and small world properties. This can be explained with the unprecedented growth that has characterized every scientific field – and the corresponding publication rates – in the recent years.

This brings us to the second finding of the present Chapter. The empirical evidence suggests the existence of some invariant mechanisms, associated with domain-related specificities, giving rise to collaboration networks. Considering that our aim is to identify the minimal set of microscopic rules able to reproduce the topology of such networks, we have investigated a different set of features on real collaboration networks. Such features are more elementary and primitive than the ones previously studied, thus representing more suitable basic blocks for our future agent-based model. The features that we have studied are: i. the size of collaboration events (i.e. firms per alliance or authors per paper), ii. the agents' *activity* (i.e. their propensity to engage in a collaboration) and iii. structural communities in the network (beyond the agents' sectoral or geographical positions).

Our findings can be summarized as follows. The distribution of agents per collaboration is broad and right-skewed for all R&D and co-authorship networks, even though the co-authorship networks exhibit a higher degree of variability across fields. The number of agents per collaboration event spans from 2 (the vast majority in all networks) to 55 (in the relativity and gravitation co-authorship network). The agents' activities distribution are dispersed and right-skewed as well, spanning several orders of magnitude. Differently from many networks indicators, the activities are stable and can effectively model the propensity of every agent to engage in a collaboration event, thus making them viable candidates for an agent attribute in our model.

Finally, we have detected the presence of modular structures in all collaboration networks, through a well-known community detection algorithm (Infomap). Such finding is significant and robust across domains. However, being it a complex and emerging topological property of the network, we decide to use it as a validity test, and not as a building block of our model. We argue that the modular structures are indicative of some specific microscopic rules of strategic link formation, that involve the presence of an agent membership attribute. The existence of this attribute, together with some rules of propagation during the establishment of collaborations, will instead be at the basis of our model; more details will follow in Chapter 4.

In conclusion, we have identified a set of three fundamental attributes that exhibit similar properties in several collaboration networks and will constitute the building blocks of an agent-based model, aimed at reproducing more sophisticated (both macroscopic and microscopic) network features. Such attributes are – in order – a broad and right-skewed distribution of agents per collaboration event, a broad and right-skewed distribution of activities (the agents' propensities to be involved in a collaboration event), and the presence of a membership attribute, that can be propagated between agents in a collaboration event and defines the network clusters.

The remaining network properties, together with the stylized facts that we have studied in Chapter 2, rather than being a starting point, will instead be used to validate our model and fine-tune its parameters in different collaboration networks. The development of the agent-based model and its validation in both the R&D and the co-authorship domains is the topic of Chapter 4.

Chapter 4

Modeling the formation of collaboration networks

Summary

In this Chapter we develop an agent based model of strategic link formation, inspired by our empirical findings on R&D and co-authorship networks. We have found that the growth of collaboration networks is driven by mechanisms which are both endogenous to the system (that is, depending on existing alliances patterns) and exogenous (that is, driven by an exploratory search for new collaborations). In order to investigate the effects and the interdependencies between these two mechanisms, we develop a general modeling framework that includes both of them and allows to tune their relative importance in the formation of links. We first test our model against the SDC Platinum alliance dataset, and then extend our validation on a large set of sectoral R&D networks as well as co-authorship networks. Remarkably, by fitting only three *macroscopic* properties of the network, our model is able to reproduce a number of *microscopic* measures characterizing the network topology, including the distributions of degree, local clustering, path length and component size, and the emergence of network clusters. Furthermore, by estimating the link probabilities, we find that endogenous mechanisms are predominant over the exogenous ones in the network formation, in most of the collaboration networks we investigate. Our framework not only brings additional support, but also precisely quantifies the importance of existing network structures for selecting collaboration partners in different domains.

Based on M. V. Tomasello, N. Perra, C. J. Tessone, M. Karsai, F. Schweitzer, “The Role of Endogenous and Exogenous Mechanisms in the Formation of R&D Networks”, *Scientific Reports*, 4, 5679 (2014). Only the model development and validation parts of the article are included in the present Chapter. M.V.T. contributed to designing the microscopic rules of the model, made the major effort in running the computer simulations, analyzed all the results and produced all the plots. The validation of the model on co-authorship networks, not included in the original article, was designed and written by M.V.T. for the present dissertation.

4.1 Modeling the growth of R&D networks

In the previous Chapters, we have systematically analyzed the salient features of R&D networks and co-authorship networks. Our results suggest the presence of a significant level of similarities in the topology of these collaboration networks and the way they evolve over time, albeit associated with domain- or sector-related peculiarities.

In particular, for the case of R&D networks, we find that two kinds of mechanisms are crucial in the formation of new alliances, in agreement with previous studies (Rosenkopf and Padula, 2008). Such mechanisms can be *endogenous* (i.e. previous alliances and previous network structures) and *exogenous* (i.e. exploratory search of new partners), with respect to the network. However, both empirical and theoretical studies have mainly focused only on one of the two mechanisms, also called "network endogeneity"(Garas *et al.*, 2014; Gulati and Gargiulo, 1999; Powell *et al.*, 1996; Walker *et al.*, 1997) and "exogenous partner selection"(Burt, 1992; Cowan and Jonard, 2004; Rosenkopf and Nerkar, 2001) respectively.

The goal of the model we develop in the present Chapter is to unify these two classes of mechanisms and quantify their relative importance in the formation of a collaboration network. We aim as well at extending the validity of such concept from R&D to co-authorship networks. Our model is intended to reproduce the main global properties and a set of microscopic measures (including degree, local clustering and path length distributions) of real networks. To this purpose, we test the model against the SDC alliance dataset, as well as the APS dataset for co-authorship networks in science. The validation of the model and the tuning of its parameters will give us insights into the micro-level decisions operated by the agents and – consequently – the growth of the networks themselves, pointing out possible similarities and differences across sectors and domains.

4.1.1 Deriving the microscopic rules for link formation

Typically, the concept of endogenous and exogenous mechanisms has been used in the management literature with respect to the belonging of firms to the R&D network. We follow such definition and refer to an alliance involving a partner that is already part of the R&D network as "endogenous". Likewise, an alliance involving a partner that is not part of the R&D network yet is referred to as "exogenous". While the endogenous mechanisms depend on the firms' social capital (describing their position in the network), the exogenous mechanisms are affected by the firms' technological and commercial capital. A firm's social capital can be further explained by two variables(Gulati, 1995b; Podolny, 1993): its *prominence* – i.e. the history of its previous alliances – and its *cohesiveness*, defined as the set of its direct and indirect links with other firms in the network. In this regard, some empirical studies(Powell *et al.*, 1996; Rosenkopf and Padula, 2008) found that several firm "clusters" populate the R&D network, thus giving rise to different kinds of alliances depending on the firms' position in the network.

In particular, three categories of R&D alliances have been identified: i. within-cluster alliances (the partners belong to the same cluster); ii. semi-distant alliances (the partners form a so-called "shortcut" between two different clusters); iii. distant alliances (at least one of the partners is an isolated node, i.e. a newcomer firm). Obviously, a certain number of R&D alliances is not explained by the partners' social capital – think, for instance, of alliances involving start-up companies or financial institutions that have no previous experience in R&D activities. One rationale for the search of this kind of partners, whose technological and commercial capital plays a crucial role, is that they can provide access to new information or unique technical knowledge.

However, neither the network endogeneity nor the exogenous partner selection, taken independently, are able to explain the topology of observed R&D networks. Endogenous mechanisms alone would lead to more and more centralized network structures over time, which we do not observe in reality (see Chapter 2). On the other hand, exogenous mechanisms alone would lead to more regular networks topologies, which we do not observe neither. There exists a prominent modeling work (Guimera *et al.*, 2005) that analyzes the formation of teams as a function of some microscopic parameters, including the team size and the propensities to select newcomers or repeat collaborations. We extend this study by considering more fine-tuned linking probabilities, adding a heterogeneous agent propensity to initiate alliances, and validating such a model for the first time, to the best of our knowledge, on a set of large inter-firm networks – and not networks of individuals.

Inspired by these considerations and by the empirical evidence, we introduce below the microscopic rules of our agent-based model.

4.1.2 Development of the agent-based model

To quickly recap, the empirical observations on collaboration networks indicate clear heterogeneities in activity and connectivity patterns, small-world features, a moderate level of transitivity, and a highly modular structure. Starting from this evidence, we consider a network composed of N nodes; each of them is endowed with two fundamental attributes, an *activity* and a *label*. Such attributes define the nodes' interaction rules, which are organized in five distinct phases, as described below.

Node activation. We assign to each of the $i = 1, \ldots, N$ nodes an activity a_i, analogous to the empirical activities extracted from the SDC dataset. Indeed, we sample without replacement all the values a_i from the empirical activity distribution. The activities we assign to the N nodes are computed by considering the entire observation period (therefore, in the case of the pooled R&D network, $a_i \equiv a_{i,t=2009}^{\Delta t=26\text{years}}$). Given the strong robustness of empirical activities to the time window, we decide to use the longest possible window, because it contains complete information about the dataset and gives activities a_i that are always strictly greater than 0 – all firms listed in the SDC dataset, by definition, must be involved in at least 1 alliance, and all authors must have co-authored at least one paper to be reported in the co-authorship network. The activity defines the propensity of each node to be involved in a collaboration event. We use this quantity to model the activation probability of each agent. In particular, at every time step, a node i initiates an alliance with probability $p_i = \eta a_i \Delta t$, and the number of active nodes N_A is

$$N_A = \eta \langle a \rangle N \Delta t, \tag{4.1}$$

where $\langle a \rangle$ is the average node activity and η is a rescaling factor that allows to adjust the activation rates, and consequently the number of active nodes per time step. We find that the model is strongly robust to the choice of η, showing no measurable changes for η ranging from 10^{-5} to 1; however, we fix η to obtain N_A roughly equal to the number of alliance events or co-authored papers per day actually reported in the datasets that we analyze. In the case of the pooled R&D network, $\eta = 0.01$. Without loss of generality, we fix $\Delta t = 1$.

Selection of the alliance size. When a node gets activated, it selects the number of partners m with whom the alliance is formed. We assume that the value of m is totally independent of any characteristic of the active node: we sample it, without replacement, from the empirical distribution of number of partners per alliance. In other words, we shuffle the sequence of number of partners per alliance, or authors per paper (directly measured from the datasets), and then extract a value every time an activation event occurs; m can be thought of as the number of agents involved in every collaboration event, diminished by 1, because the active node is not counted twice.

Label propagation. As we have previously shown in Chapter 3, real collaboration networks are organized in clusters of tightly interconnected nodes. However, these clusters are not isolated; in the case of R&D networks, previous studies (Rosenkopf and Padula, 2008) have detected the existence of "shortcuts" connecting different clusters, as well as the formation of alliances with new partners not yet belonging to the R&D network. This observation suggests that firms diversify some of their alliances, rather than just establishing collaborations within a specific cluster. We model this feature assuming that each of the N nodes is endowed with an attribute named *label*. This attribute is unique – i.e. every node can have only one label at any time – and fixed – once a node assumes a label, it does not change –. Labels model the belonging of the firms to different groups that they implicitly define with their shared practices and commonly recognized behaviors: in other words, a label symbolizes the membership of the firm in a well defined and recognized "club" or "circle of influence". In addition, we assume that such membership can be transferred to other firms as a consequence of an alliance, provided that they are not part of any circle of influence yet.

The hypothesis of such a membership attribute is in agreement with the results reported by Yang and Leskovec (2012), that have identified the presence of communities based on ground truth in real networks. Such communities include nodes that do not necessarily share features such as the same geographical provenience, or the belonging to the same institution. They rather define these communities dynamically, through consecutive interactions and link formations, phenomenon that is captured by our membership attribute and its propagation. We argue that the same reasoning holds for co-authorship networks, where clusters of collaborating authors are formed depending not only on their geographical or scientific distance, but also through subsequent propagation of an implicit membership attribute. In our network representation, every collaboration initiator does indeed propagate its label to all of its m partners, if they are non-labeled. At the beginning of every simulation, all nodes are *non-labeled*, meaning that their membership attribute is blank. There are two ways a non-labeled node can assume its label: (i) the node either receives the label from another node, if the latter initiates an alliance, or (ii) it takes an arbitrary and unique label when it becomes active for the first time (see Fig. 4.1).

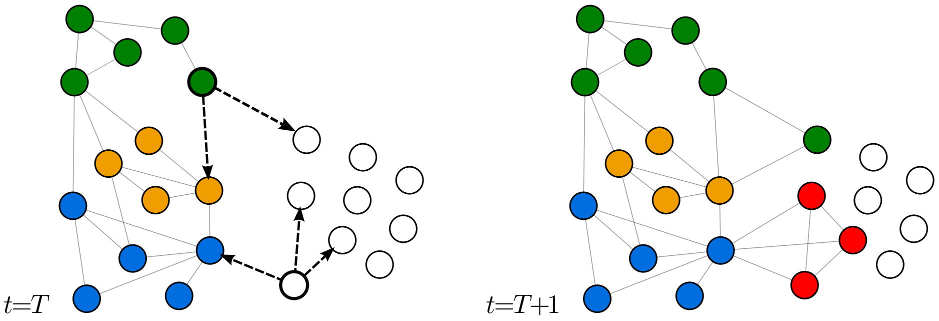


Figure 4.1: Two representative examples of label propagation. A labeled node (whose label is depicted in green) chooses to form an alliance with $m = 2$ partners, one having a different label (depicted in yellow) and one non-labeled, at time $t = T$. The initiator propagates its green label at time t=T+1 only to the previously non-labeled node. The link with the yellow node is still formed, but the label propagation does not occur. Likewise, a non-labeled node gets activated at time t=T and forms an alliance with $m = 3$ partners, two non-labeled nodes and one labeled (blue) node. The non-labeled initiator takes a new arbitrary label (depicted in red) at time t=T+1 and propagates it only to its previously non-labeled partners. The red label is not propagated to the blue node, even though the links are regularly formed.

Selection of the partner categories. The presence of node labels induces different types of alliances, that we explicitly distinguish in our model (see Fig. 4.2). This is in line with previous studies on team assembly mechanisms (Guimera *et al.*, 2005), that we extend to collaboration networks whose agents can be represented also by firms, and not only individual scientists or artists. In particular, we assume that if the alliance initiator is a labeled node, it represents an *incumbent* firm, i.e. a firm that has already been involved in at least one alliance, or an *expert* author, i.e. an author that has co-authored at least one scientific paper. In this case, the initiator can link to a labeled node having the same label (with probability p_s^L), or to a node having a different label (p_d^L), or to a node without label (p_n^L). If the initiator is a non-labeled node, it represents a *newcomer* firm, i.e. a firm that has not been involved in any alliance event yet, or a *novice* scientific author. In this case, the initiator can link to a labeled node (with probability p_l^{NL}), or to another non-labeled node (p_{nl}^{NL}). The five probabilities associated to these occurrences, represented in Fig. 4.2, are the free parameters of our model.

Following the definitions traditionally adopted in previous theoretical literature, we argue that the probabilities associated to a connection with a labeled node (p_s^L, p_d^L and p_l^{NL}) quantify the relevance of endogenous mechanisms for link formation, given that the initiator of the alliance has information about the network position (i.e. social capital) of its potential partners. Likewise, the probabilities associated to a connection with a non-labeled node (p_n^L and p_{nl}^{NL}) estimate the relevance of the exogenous mechanisms. In this case, the initiator cannot have any information about the social capital of a firm (or an author) that is not part of the network yet; the choice to initiate a collaboration event is due to different rationales, such as the technological, scientific or geographical proximity of the agents. It should be noted that the agents in this phase select the *category* of their partners, not the specific partners themselves. The selection of such categories is made independently of their population: this means that the initiator only selects the pool in which the potential partner will be, and only afterwards the actual link will be formed. Obviously, the population of every category and every circle of influence changes dynamically as the network evolves. At the beginning of the network evolution, for instance, all nodes are non-labeled and the few existing circles of influence slowly grow around the few active nodes. This kind of dynamics is not only in agreement with some theoretical arguments (Kahneman and Tversky, 1996), but also capable of originating the features that we observe in real collaboration networks, as we show below. The five probabilities are bounded by two conditions, reducing the number of independent parameters to three; their nomenclature and their meaning are summarized in Table 4.1.

Link formation. After deciding the category of each of its m partners, we assume that the initiator selects its specific partners within those categories according to their attractiveness. Indeed, it has been shown for the case of R&D networks (Gulati, 1995b; Podolny,

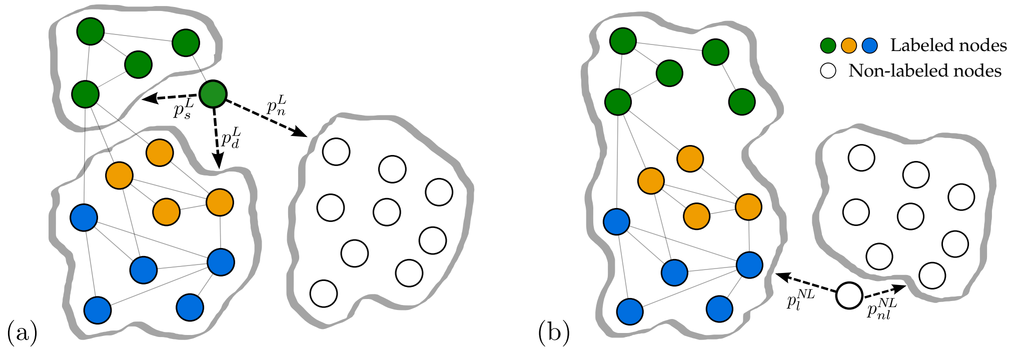


Figure 4.2: Selection of partner categories. (a) If a labeled node (depicted in green) gets activated, it has 3 choices: it can link to a labeled node having the same label with probability p_s^L, or to a labeled node having a different label with probability p_d^L, or to a non-labeled node with probability p_n^L. (b) If a non-labeled node (depicted in white) gets activated, it has 2 choices: it can link to a labeled node with probability p_l^{NL}, or to another non-labeled node with probability p_{nl}^{NL}.

Parameter	Meaning	Type of mechanism
$\boldsymbol{p_s^L}$	**Probability of a labeled node to select a node with the same label**	**Endogenous**
$\boldsymbol{p_d^L}$	**Probability of a labeled node to select a node with a different label**	**Endogenous**
p_n^L	Probability of a labeled node to select a non-labeled node	Exogenous
$\boldsymbol{p_{nl}^{NL}}$	**Probability of a non-labeled node to select a non-labeled node**	**Exogenous**
p_l^{NL}	Probability of a non-labeled node to select a labeled node	Endogenous

Table 4.1: Model parameters and their explanation. We have two binding conditions, reducing the number of independent parameters to three: the probabilities p_s^L, p_d^L and p_n^L sum up to 1. Likewise, p_{nl}^{NL} and p_l^{NL} sum up to 1. We report the probabilities that we choose as independent parameters in bold character.

1993) that alliances tend to be directed towards firms having a higher prominence (i.e. history of previous alliances). We argue that the same holds for co-authorship networks, and model this by considering the degree of each potential partner. More precisely, we use a linear preferential attachment rule, where the probability to attach to a node j linearly scales with its degree k_j, meaning that $\Pi(k_j) \sim k_j$. The preferential attachment rule is applied within the pool of all candidate partners, once the selection of the partner category has been made by the collaboration initiator (see Fig. 4.2). This rule obviously does not apply when the initiator – be it labeled or not – decides to connect to a non-labeled node, which has by definition no previous partners ($k_j = 0$). In this case, the partner is selected among all non-labeled nodes with equal probability. When the selection process is complete, the initiator connects to its m partners. In agreement with our representation of the R&D and co-authorship networks, we assume that all the m partners will also link to each other, forming a fully connected clique of size $m + 1$.

4.2 Model validation

For the sake of clarity, in the current Section we describe the validation procedure of our model against one dataset, namely the SDC Platinum alliance database. The validation and fine tuning of the parameters will be also performed on other datasets – as we explain in Section 4.3 – by adjusting a set of relevant measures, but keeping the same exact procedure.

4.2.1 Implementation and parameter space exploration

We perform extensive computer simulations by applying the microscopic rules described in Section 4.1.1, and varying the values of the independent parameters. We fix the model parameters that we can directly measure from the data, namely the number of agents (in this case, $N = 14,561$), the distribution of the node activities a_i, and the distribution of number of partners m per alliance event. We stop every computer simulation when the total number of formed alliances equals the number of alliance events reported in the dataset (in this case, $E = 14,829$).

We vary the values of p_s^L, p_d^L and p_{nl}^{NL} in discrete steps spaced by 0.05, in the interval $(0, 1)$. The parameters p_s^L and p_d^L are bounded by the condition $p_n^L = 1 - p_s^L - p_d^L \geq 0$, meaning that their sum has to be smaller or equal to 1. This condition translates into 3,249 points to explore in the 3-dimensional parameter space, for each of which we run 200 simulations (for a total of 649,800 runs). Similarly to the previously analyzed collaboration networks (see Chapter 3), we test the final aggregated network resulting from each of the 649,800 computer simulations with respect to three properties: average degree $\langle k \rangle$, average path length $\langle l \rangle$ and global clustering coefficient C. Remarkably, we find that all such quantities are distributed around the real values for all the collaboration networks we study, as we report in Appendix C more detail. This testifies that our model well captures the topology of many observed networks for a large set of free parameters.

It should be noted that we have imposed a few features from the empirical networks (number of nodes N and alliances E, and the distributions of node activities a_i and partners per alliance m). However, the distributions of the simulated $\langle k \rangle$, $\langle l \rangle$ and C obtained by exploring the parameter space of the model, although centered around the real values, exhibit a fairly large variance (as reported in Appendix C), thus allowing a meaningful exploration of the parameter space. As a logical consequence, we aim at identifying which parameter combination is able to give the best match with the real network. To this purpose, we use a Maximum Likelihood approach. The peculiarity of this study is that, instead of having a set of observations against which we can validate our model, we only have one empirical point: the real network. In particular, we cannot consider the three measures as independent, therefore the Likelihood function $\mathcal{L}$ reads as:

$$\mathcal{L}(p|net^{OBS}) = f(net^{OBS}|p) \tag{4.2}$$

where $f(\cdot)$ is the joint density function of all parameter combinations p resulting in a network that is equivalent to the observed one net^{OBS}. Both p and net^{OBS} are vectors with three components, expressing respectively the three model parameters $p \equiv (p_s^L, p_d^L, p_{nl}^{NL})$ and the three global network measures $net^{OBS} \equiv \big(\langle k\rangle^{OBS}, \langle l\rangle^{OBS}, C^{OBS}\big)$. Therefore, we need to find the parameter combination $(p_s^L, p_d^L, p_{nl}^{NL})$ maximizing the Likelihood $\mathcal{L}(p|net^{OBS})$ to generate a network whose macroscopic properties are *sufficiently similar* to the real network net^{OBS}. By this, we mean that the relative errors from the observed values for the average degree $\varepsilon_{\langle k\rangle}$, the average path length $\varepsilon_{\langle l\rangle}$ and the global clustering coefficient ε_C have to be smaller than a certain threshold ε^0.

We empirically compute the Likelihood function $\mathcal{L}$ for each point in the parameter space by counting the fraction of its 200 simulation realizations that fulfill the criteria $\varepsilon_{\langle k\rangle} < \varepsilon^0$; $\varepsilon_{\langle l\rangle} < \varepsilon^0$; $\varepsilon_C < \varepsilon^0$. This way, we obtain values that can range from 0 (no realization of that parameter combination fulfills the criteria) to 1 (all of its realizations fulfill the criteria).

The error threshold value ε^0 we impose for the computation of the Likelihood score influences the number of points in the parameter space that fulfill our matching criteria. Obviously, by decreasing ε^0, we observe a smaller number of points displaying high likelihood scores, as we could expect, because a better representation of reality is required (see Appendix C). We take a conservative approach and use an error threshold $\varepsilon^0 = 0.02$, that ensures a good matching with the observed R&D network, without cutting out too many points in the parameters space.

The corresponding Likelihood scores are reported in Fig. 4.3 by means of a 3-dimensional color map, where the color scale is representative of the Likelihood. To have a more detailed representation of the likelihood scores, we also show three slices of the parameter space obtained by fixing the parameter p_s^L in the range $0.25 \div 0.35$, corresponding to the highest likelihood score region, always using the error threshold $\varepsilon^0 = 0.02$. The 2-dimensional color maps reported in Fig. 4.3 depict the likelihood score as a function of the other two free parameters p_d^L and p_{nl}^{NL}.

The point with the highest likelihood score, for the pooled R&D network, has the following coordinates in the parameter space: $p_s^{*L} = 0.3$, $p_d^{*L} = 0.3$ and $p_{nl}^{*NL} = 0.25$. We can already see that in the optimal configuration, labeled nodes exhibit a balanced alliance strategy, with $p_s^{*L} = 0.3$, $p_d^{*L} = 0.3$, and consequently $p_n^{*L} = 0.4$, while the non-labeled nodes exhibit a strong tendency to connect to labeled nodes ($p_l^{*NL} = 0.75$), as opposed to a low linking probability with other non-labeled nodes ($p_{nl}^{*NL} = 0.25$). This tendency, as we show later, is common to most collaboration networks.

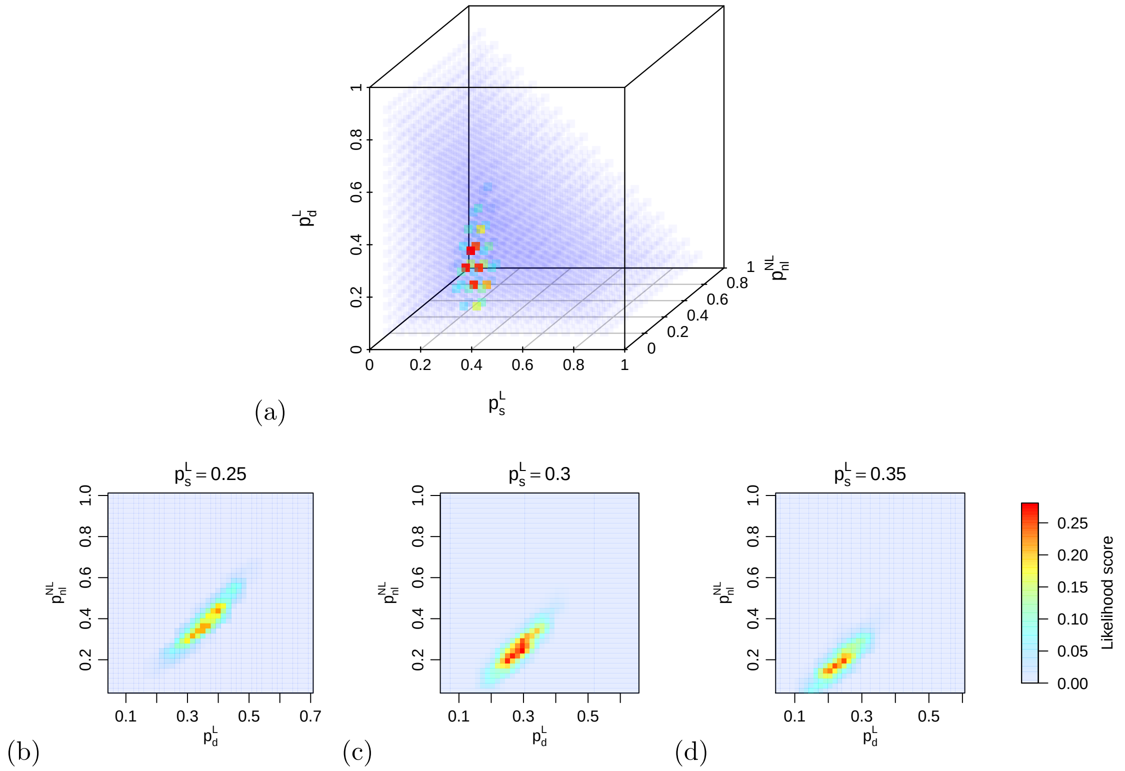


Figure 4.3: Likelihood scores for all points in the parameter space, for $\varepsilon^0 = 2\%$, represented with a 3-dimensional color map (a). After fixing the value of p_s^L to 0.25 (b), 0.3 (c) and 0.35 (d), we report the Likelihood score as a function of p_d^L and p_{nl}^{NL}, using the same color scale.

Optimal simulated R&D network				**Observed R&D network**	
Model parameter	Value	Measure	Value	Measure	Value
p_s^{*L}	0.3	$\langle k \rangle^*$	2.764 ± 0.006	$\langle k \rangle^{OBS}$	2.736
p_d^{*L}	0.3	$\langle l \rangle^*$	5.329 ± 0.068	$\langle l \rangle^{OBS}$	5.412
p_n^{*L}	0.4	C^*	0.098 ± 0.005	C^{OBS}	0.101
p_{nl}^{*NL}	0.25				
p_l^{*NL}	0.75				

Table 4.2: Model parameter set p^* defining the optimal simulated R&D network. The average degree, average path length and global clustering coefficient of the 200 realizations of the optimal R&D network are compared to their analogous empirical values.

We report in Table 4.2 the set of parameter values maximizing the likelihood score, together with the values of average degree, average path length and global clustering coefficient for the simulated and the real R&D networks. From now on, we call the network generated with these parameters the *optimal simulated R&D network*. More precisely, we generate 200 realizations of the optimal simulated R&D network (as well as of any other network with a generic parameter set). For this reason, the results we present in the next section are computed on all the 200 realizations of such optimal network.

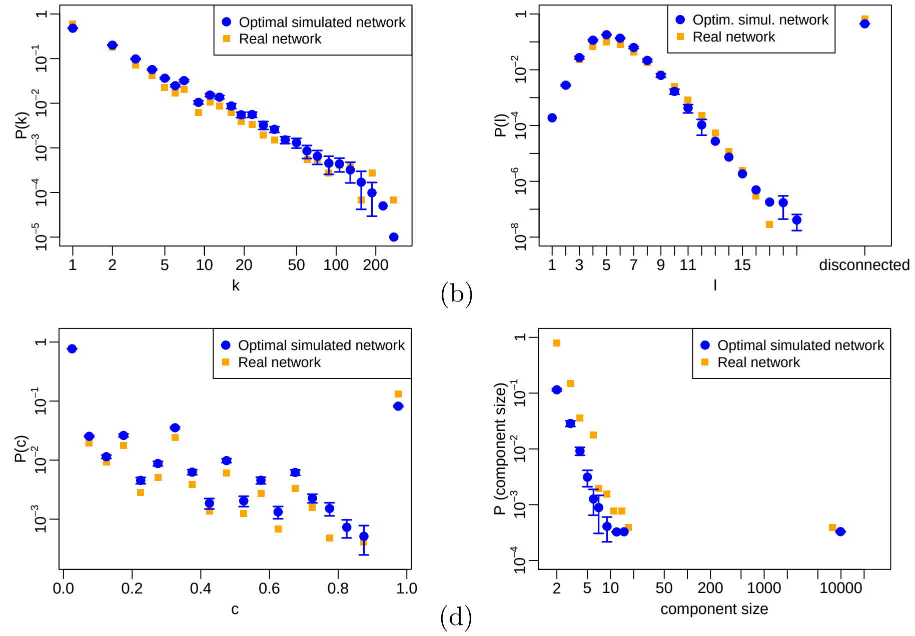
The optimal set of linking probabilities gives some interesting insights into the nature of the strategies pursued by the agents in a collaboration network. In the case of the pooled R&D network, we find that all firms tend to have a preference to link incumbent firms: 60% of the alliances initiated by incumbents belong to this category ($p_s^{*L} + p_d^{*L}$), as well as 75% of the alliances initiated by newcomers p_l^{*NL}. This result in is line with well-known economic theories(Gulati, 1995a) that have shown how previous interactions between two firms increase the likelihood of future alliances among them if they are already part of the R&D network. In addition, newcomers are incentivated to join the R&D network by partnering with firms that are already part of it(Ahuja, 2000b). On the other hand, we find that 40% of the alliances initiated by incumbents, as well as 25% of the alliances initiated by newcomers, are directed to newcomers. These alliances can be driven only by exogenous factors(Rosenkopf and Nerkar, 2001), and a possible explanation behind this tendency is the appealing of newcomers' commercial or technological capital.

Overall, our findings suggest that both endogenous and exogenous mechanisms contribute to the alliance formation. However, the first class appears to be more prominent: the fine tuning of our model provides additional evidence, and a precise quantification, of how previous network structures play the biggest role in deciding the potential partners when a new alliance is formed. As reported in the literature(Gulati, 1995a; Podolny, 1993), the belonging to the R&D network, and in particular the belonging to a specific circle of influence, signals a firm's reliability and competencies to potential partners. This mechanism is clearly predominant over the exogenous search for alliance partners, hence we argue that being aware of the partners' positions in the R&D network is of fundamental importance for every firm.

4.2.2 Validation and further tests

The optimal parameter set for the case of the pooled R&D network, as we have shown above, is $p^* \equiv (p_s^{*L} = 0.3;\ p_d^{*L} = 0.3;\ p_n^{*L} = 0.4;\ p_{nl}^{*NL} = 0.25;\ p_l^{*NL} = 0.75)$. As a second, last step in our validation procedure, we now want to check whether our model, fed with these parameter values, is able to reproduce further microscopic properties of the real network. To this purpose, we report in Fig. 4.4 four additional distributions computed on the optimal simulated R&D network – node degrees, path lengths, local clustering coefficients and component sizes – and compare them to the empirical ones. From now on, in every plot we show, the blue circles correspond to the mean values and the error bars correspond to the standard deviations of all the quantities we analyze on the 200 realizations of the optimal simulated R&D network.

Remarkably, we find that our model is able to reproduce all the distributions, namely the typical right-skewed degree distribution, the path length distribution peaked around the mean value 5 and the local clustering coefficient distribution. The model can also



(a) (b) (c) (d)

Figure 4.4: Distributions of node degrees (a), path lengths (b), local clustering coefficients (c) and component sizes (d) for the real and the optimal simulated networks. Most of the error bars are not visible, because the values are very narrowly distributed across the 200 realizations of the optimal simulated network.

reproduce the component size distribution, showing the emergence of a giant component in the network (containing roughly 60% of the nodes) together with many smaller components down to size two. Isolated nodes (nodes with degree equal to 0) are excluded from our representation; hence, the smallest observable component size in our networks is 2.

Going further in our validation procedure, we test the modular properties of the optimal simulated R&D network. As already done in Chapter 3, we run the Infomap community detection algorithm (Rosvall and Bergstrom, 2008) on all of the network realizations. We identify the presence of 1,600 ± 20 clusters (whilst 3,500 clusters populate the empirical R&D network), whose minimum size is 2 and maximum size is around 100 nodes, similarly to the empirical network (see Fig. 3.14). We report in Fig. 4.5 a visual representation of the optimal simulated R&D network and the size distribution of the detected clusters.

Interestingly, this distribution resembles the one of the empirical R&D network, with the only exception of having significantly fewer counts related to small clusters of size 2 and 3. The larger clusters, up to 100 nodes, that dominate the network structure and contribute to its modularity, are equally populating the empirical and the optimal simulated R&D networks. Another evidence of their similarity is the modularity score of the optimal simulated R&D network $Q^* = 0.66 \pm 0.01$, surprisingly close to its empirical analogue $Q^{OBS} = 0.68$. Also in the case of the optimal simulated R&D network, its modularity score Q^* is significantly greater (with a p-value computationally indistinguishable from zero) than the ones obtained for a set of 500 randomly generated networks with the same degree sequence (whose Q is normally distributed around 0.485 with a standard deviation of 0.001), showing that the modularity is not an artifact of the network size and density.

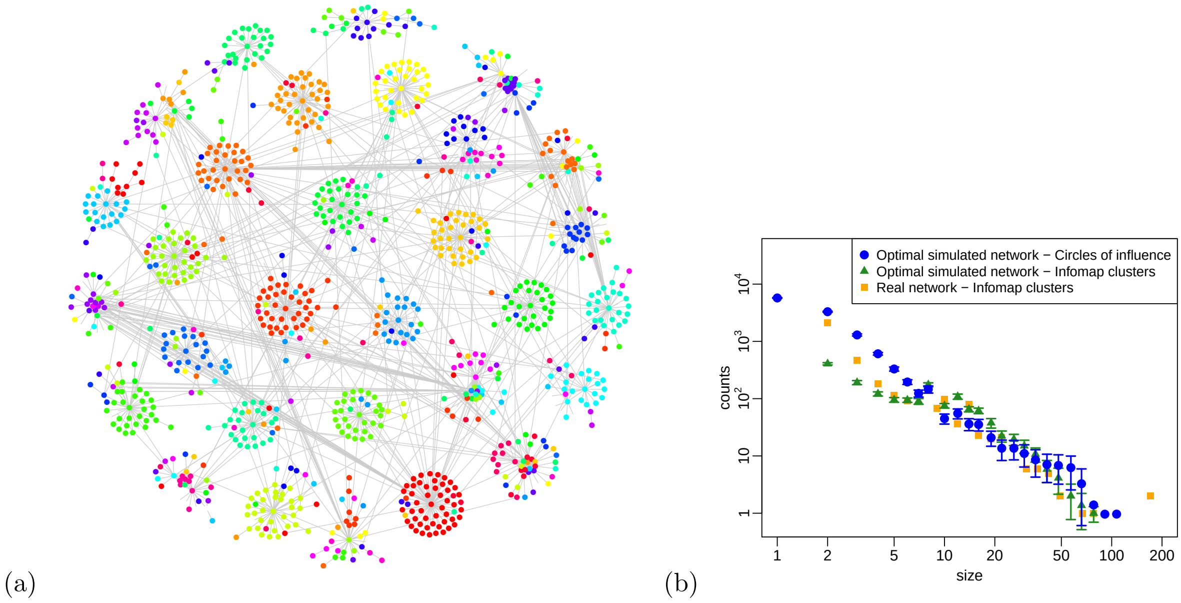


Figure 4.5: (a) Visual representation of one realization of the optimal simulated R&D network, using the Fruchterman-Reingold layout(Fruchterman and Reingold, 1991) and considering only the 30 largest clusters detected by the Infomap algorithm. Distinct clusters are represented by grouping nodes in distinct regions of the plot area. In addition, we depict our node labels by using different colors; it is clearly observable that most of the nodes in a given cluster share the same label. (b) Size distribution of i. the circles of influence in the 200 realizations of the optimal simulated R&D network, ii. the Infomap clusters in the 200 realizations of the optimal simulated R&D network and iii. the Infomap clusters in the empirical R&D network.

We now test whether our node labels are actually able to reproduce such a modular structure of the network. In order to estimate the overlap between the clusters detected via the Infomap algorithm and the circles of influence defined by our node labels, we compute the *normalized mutual information* coefficient I_{norm}(Danon *et al.*, 2005), very often used to this purpose(Lancichinetti and Fortunato, 2009). Given two network partitions A and B, the value of the coefficient $I_{\text{norm}}(A, B)$ ranges from 0 (if the partitions A and B are independent) to 1 (if the partitions A and B are identical). In our case, we obtain a striking $I_{\text{norm}}(\text{Labels, Infomap clusters}) = 0.899 \pm 0.001$, testifying how well our node labels capture the emergence of clusters in the R&D network. We also present a visual comparison of the clusters identified by means of Infomap with the circles of influence resulting from the implementation of our model in Fig. 4.5. Similarly to the empirical R&D network, we consider only the 30 largest Infomap clusters in the optimal simulated R&D network and visualize them by grouping the corresponding nodes in distinct regions of the plot; in addition, here we depict our node labels with arbitrary colors. As testified by the high normalized mutual information score, our visual example nicely confirms that most of the nodes in a given cluster share the same label. The size distribution of the circles of influence defined by these labels is also shown in Fig. 4.5. Its similarity to the size distribution of the Infomap clusters in both the empirical and the optimal simulated R&D network provides another evidence of the goodness of our model.

In order to estimate to what extent our link formation rules capture the decision making process made by real firms, we test the optimal simulated network with respect to a feature that is both *microscopic* and *dynamic*: the distribution of path lengths between every pair of nodes at the moment of the link formation. This should not be confused with the path lengths analyzed before, whose distribution was computed on the final aggregated R&D network, between *every* pair of nodes, in both the real and the simulated case. Now we only consider pair of nodes that eventually form a link between each other. More precisely, we plot the distribution of the path lengths between two firms as of the day before their alliance formation (for the real R&D network) and the path lengths between two nodes at the time step preceding the link formation (for the optimal simulated R&D network). We also consider as separated counts all those alliance events involving at least one newcomer firm (or an isolated node, in the simulated network).

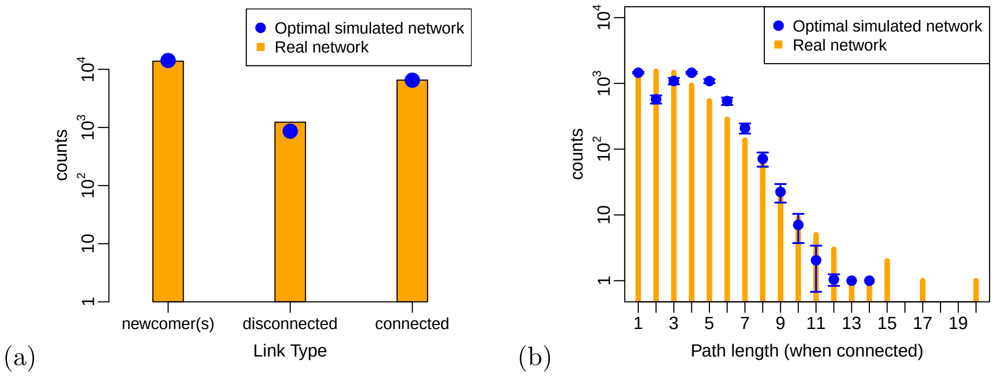


Figure 4.6: (a) Distribution of link types for the real and the simulated R&D networks. “Connected” refers to nodes already belonging to the same connected component of the network prior to the link formation; “disconnected” refers to nodes already belonging to the network, but placed in two disconnected components; “newcomer(s)” means that at least one of the nodes was isolated (i.e. not yet part of the network) before the link formation. (b) Distribution of path lengths at the moment of link formation (only for nodes belonging to the same connected component).

We show our findings in Fig. 4.6. The model can reproduce the counts of links formed between (i) firms belonging to the same connected component of the network, (ii) firms belonging to different disconnected components and (iii) involving at least one newcomer (isolated) firm. Furthermore, the model reproduces also the counts relative to nodes which are already connected by a path before the link formation. The only small discrepancies can be observed in correspondence to path lengths equal to 2 and 3, due to effects of triadic and cyclic closure exhibited by real firms that are not fully captured by our model, as we already anticipated. However, our model correctly predicts the formation of links between nodes that are relatively distant in the network or even already directly connected – the cases when the path length is equal to 1 are related to the same two partners engaging in a new alliance.

In conclusion, we find that our model, although tuned only considering three global static measures, provides a surprisingly good prediction of several microscopic and dynamic features, such as the distributions of degree, local clustering, path length and component size, the emergence of network clusters and, even more remarkably, the distribution of path lengths at the moment of the alliance formation.

4.3 Comparing different collaboration networks

In the present Section, we extend the validation of our model to a larger set of collaboration networks. We adopt the same procedure described in Section 4.2 and replicate it on the different datasets, by appropriately adjusting the relevant empirical parameters – namely, the number of nodes N and collaboration events E, and the distributions of node activities a_i and partners per collaboration m. The datasets against which we test our model represent six sectoral R&D networks, extracted from the SDC dataset through the procedure described in Chapter 4, and six co-authorship networks, extracted from the APS dataset (see Chapter 3).

The six R&D networks are related to the sectors of computer software, pharmaceuticals, R&D laboratory and testing, computer hardware, electronic components and communications equipment. The six co-authorship networks are instead related to the scientific fields of quantum mechanics, field theories, and special relativity; general relativity and gravitation; optics; electronic transport in condensed matter; superconductivity; other areas of applied and interdisciplinary physics (this field includes network theory). We list the main empirical properties of these networks in Table 4.3, including also the pooled R&D network for comparison.

The quantities reported in Table 4.3 are used to calibrate our model, together with the distributions of agent activities a_i and number of partners m per collaboration event. For every collaboration network, we then explore the parameter space, in order to find the probability set ensuring the best match with the corresponding empirical network. This procedure requires a remarkable computational effort; each of the 12 collaboration networks originates a parameter space composed of 3,249 points, for each of which we run 25 computer simulations – for a total of around 1 million simulations. Even though the realizations of each parameter set that we obtain with such computationally cumbersome procedure are 25, we find that all of them exhibit network properties with a very small

	N	E	**Links**	$\langle k\rangle^{\mathrm{OBS}}$	$\langle l\rangle^{\mathrm{OBS}}$	C^{OBS}
Pooled R&D network	14,561	14,829	21,572	2.74	5.41	0.101
Sectoral R&D networks						
Pharmaceuticals (SIC 283)	3,829	5,277	6,019	3.14	4.94	0.097
Computer hardware (SIC 357)	1,582	2,672	4,047	5.12	3.70	0.161
Communications equipment (SIC 366)	1,133	1,888	2,726	4.81	3.75	0.203
Electronic components (SIC 367)	1,615	2,574	3,756	4.65	3.80	0.168
Computer software (SIC 737)	3,381	4,134	5,862	3.47	4.33	0.138
R&D, laboratory and testing (SIC 873)	3,188	4,032	5,364	3.37	5.15	0.205
Co-authorship networks						
Quantum mechanics, field theories, special relativity (PACS 03)	21,501	19,647	56,111	5.22	6.43	0.379
General relativity and gravitation (PACS 04)	8,294	8,158	32,513	7.84	6.27	0.666
Optics (PACS 42)	27,436	20,105	94,961	6.92	6.40	0.425
Electronic transport in condensed matter (PACS 72)	19,492	11,687	55,818	5.73	7.06	0.448
Superconductivity (PACS 74)	14,920	10,541	52,615	7.05	5.87	0.443
Other areas of applied and interdisciplinary physics (PACS 89)	4,881	2,873	8,777	3.60	8.28	0.462

Table 4.3: Main empirical properties for the entire set of studied collaboration networks. For each network, we report the number of nodes N, of collaboration events E, of resulting links in our network representation, and the observed values of average degree $\langle k\rangle^{\mathrm{OBS}}$, average path length $\langle l\rangle^{\mathrm{OBS}}$ and global clustering coefficient C^{OBS}.

variance – similarly to the pooled R&D network (see Section 4.2) – and obtaining additional statistics would not significantly improve our results.

We report the optimal parameter set for every collaboration network in Table 4.4, together with the resulting mean values of average degree, path length and clustering coefficient for the optimal simulated networks. It should be noted that – given the extreme variability of the networks we test, in terms of size, density and modularity – we are forced to adjust the error threshold value ε^0 (see Section 4.2), in order to find a meaningful number of parameter configurations that are able to reproduce the empirical network with a precision ε^0. In particular for some co-authorship networks, we are not able to retrieve the average degree $\langle k\rangle$, the average path length $\langle l\rangle$ and the global clustering coefficient C with an accuracy as low as 2% (which we could achieve for the pooled R&D network). However, all the values we obtain for our simulated networks are fairly accurate and deviate from the empirical values by less than 12%, with the only exception of one co-authorship network (general relativity and gravitation).

The analysis of the optimal parameter sets reveals an even more surprising level of similarity among the collaboration networks that we have studied. First of all, the exact same model is able to reproduce the topology of all networks, in terms of average degree, average path length and global clustering coefficient, with an accuracy of at most 12%. The only exception is represented by the co-authorship network in the field of general relativity and gravitation (PACS number 04), for which the model fails to generate a network matching all the three measures $\langle k\rangle$, $\langle l\rangle$ and C at the same time. We argue that this is due to the bi-modal distribution of the partners per collaboration – or, precisely, authors per paper – in this scientific field. Differently from the other co-authorship networks, which exhibit

	ε^0	$\langle \boldsymbol{k} \rangle^*$	$\langle \boldsymbol{l} \rangle^*$	$\boldsymbol{C}^*$	$\boldsymbol{p_s^{*L}}$	$\boldsymbol{p_d^{*L}}$	$\boldsymbol{p_n^{*L}}$	$\boldsymbol{p_l^{*NL}}$	$\boldsymbol{p_{nl}^{*NL}}$
Pooled R&D network	2%	2.76	5.33	0.098	0.30	0.30	0.40	0.75	0.25
Sectoral R&D networks									
Pharmaceuticals (SIC 283)	2%	3.13	4.95	0.097	0.35	0.35	0.30	0.80	0.20
Computer hardware (SIC 357)	6%	5.37	3.59	0.175	0.55	0.30	0.15	0.90	0.10
Communications equipment (SIC 366)	2%	4.83	3.76	0.210	0.75	0.15	0.10	0.80	0.20
Electronic components (SIC 367)	2%	4.76	3.83	0.174	0.65	0.20	0.15	0.90	0.10
Computer software (SIC 737)	3%	3.56	4.27	0.141	0.55	0.20	0.25	0.95	0.05
R&D, laboratory and testing (SIC 873)	3%	3.30	5.22	0.200	0.40	0.40	0.20	0.20	0.80
Co-authorship networks									
Quant. mech., field theor., spec. relativity (PACS 03)	12%	5.83	5.58	0.392	0.85	0.05	0.10	0.45	0.55
General relativity and gravitation (PACS 04)	*> 30%**	*16.64*	*4.39*	*0.535*	*0.50*	*0.05*	*0.45*	*0.05*	*0.95*
Optics (PACS 42)	10%	7.60	5.79	0.451	0.60	0.05	0.35	0.35	0.65
Electronic transport in condensed matter (PACS 72)	8%	6.15	6.58	0.471	0.50	0.05	0.45	0.30	0.70
Superconductivity (PACS 74)	7%	7.51	5.51	0.465	0.55	0.05	0.40	0.35	0.65
Other applied and interdisciplin. physics (PACS 89)	8%	3.82	7.82	0.501	0.65	0.05	0.30	0.25	0.75

Table 4.4: Summary of all optimal simulated network statistics. For each collaboration network, we report the error threshold or accuracy ε^0; the mean values over the 25 network realizations of average degree $\langle k \rangle^*$, average path length $\langle l \rangle^*$ and global clustering coefficient C^*; the optimal linking probabilities, namely the probability of a labeled node to select a node with the same label (p_s^L), a node with a different label (p_d^L) and a non-labeled node (p_n^L), plus the probability of a non-labeled node to select a labeled node (p_l^{NL}) and a non-labeled node (p_{nl}^{NL}). The probabilities p_s^L, p_d^L and p_n^L sum up to 1; likewise, p_l^{NL} and p_{nl}^{NL} sum up to 1.
*The model is unable to generate a network matching all the three measures $\langle k \rangle$, $\langle l \rangle$ and C at the same time, for the co-authorship network in general relativity and gravitation (PACS 04). Only $\langle l \rangle$ and C can be retrieved with an accuracy of 30%, while the generated $\langle k \rangle$ is not compatible with the empirical measure. Even though we report these values for the sake of completeness, they cannot be considered significant.

broad partner number distribution with at most a heavy tail, this co-authorship network is characterized by the presence of two distinct groups of authors in the sub-fields of general relativity and gravitation, having different behaviors in terms of publications and number of co-authors (for more details, see Fig. 3.8 in Chapter 3). This bi-modality changes the intrinsic nature of this network, thus rendering the model unable to isolate and capture the two behaviors behind the link formation.

However, the model is able to reproduce two of the three measures (namely, the average path length and the global clustering coefficient) also for this collaboration network, with an accuracy of 30%. Even though the value of the average degree is not compatible with its empirical equivalent, we report in Table 4.4 the parameter set generating this network for the sake of completeness; the linking probabilities and all the other results associated to this co-authorship network, obviously, cannot be considered significant.

Similarities and differences. The optimal parameter sets show that, in all collaboration networks, nodes that are already part of the network – i.e. incumbent nodes, or labeled nodes in our network representation – tend to form links with other incumbent nodes, rather than newcomer nodes. Looking at Table 4.4, we indeed find that the probability of forming links directed from labeled nodes to other labeled nodes, that reads as $p_s^{*L} + p_d^{*L}$, is larger than 55% for all networks. Obviously, the probability of an incumbent node to choose a newcomer node as collaboration partner is always smaller than 45%.

In particular, we find that the probability of a collaboration initiator to choose partners from the same circle of influence is always greater than or equal to the probability to choose them from a different circle of influence: for all networks, $p_s^{*L} \geq p_d^{*L}$. Indeed, p_s^{*L} is strictly greater than p_d^{*L} for the totality of the co-authorship networks and almost all the sectoral R&D networks, with the exception of pharmaceuticals and R&D, laboratory and testing. These two sectors are characterized by a high technological dynamism which, we argue, is reflected on a more diversified strategy of the alliance initiators when searching for new partners. Interestingly, the same remark holds for the pooled R&D network, that includes all industrial sectors and thus combines all possible alliance strategies; this results in equal linking probabilities for labeled nodes to connect with nodes in the same or in a different circle of influence.

The tendency of collaboration initiators to select partners having the same membership attribute is much stronger in co-authorship networks. Indeed, all the examined co-authorship networks exhibit a noteworthy characteristic: the probability of an expert scientist (i.e. a person that has already authored at least one paper) to select a co-author from his/her own circle of influence is at least 10 times bigger than the probability to select a co-author from a different circle of influence (p_d^{*L} equals the lowest possible value, 5%, in all cases).

These findings support and generalize the claim that network endogenous factors are predominant in the formation of new collaborations, or – in other words – the existing network structures explain most of the newly formed links. This holds for both sectoral R&D networks and co-authorship networks.

As for the strategy of newcomer (or non-labeled) nodes, we detect some differences between the R&D and the co-authorship networks. While in the totality of co-authorship networks, newcomer nodes tend to enter the network by forming links with other newcomer nodes ($p_{nl}^{*NL} \geq 0.55$), in almost all the sectoral R&D networks, newcomers tend to enter the network by forming links with incumbent nodes ($p_l^{*NL} \geq 0.75$). The only exception is represented by the R&D, laboratory and testing sector, where the majority of links initiated by newcomers are directed towards other newcomers ($p_{nl}^{*NL} = 0.8$). The fact that $p_{nl}^{*NL} \geq 0.55$ in co-authorship networks nicely represents a typical observed behavior: when a scientist authors his/her first paper, the majority of his/her co-authors are young scientists as well (i.e. scientists that have not co-authored a paper yet), while a smaller part of the co-authors is represented by expert scientists (typically, post-docs or the professor from the same research group).

In conclusion, all the co-authorship networks, plus one R&D network in a highly technological and dynamic sector, are characterized by a tendency of newcomers to form their first links with other newcomer nodes. Conversely, in the almost totality of R&D sectors, including the pooled R&D network, newcomers tend to enter the network by forming links with incumbents.

Additional tests on collaboration networks. Similarly to the approach adopted in Section 4.2.2, we now investigate whether the model is able to reproduce a number of microscopic network properties of the collaboration networks, even without imposing them in the validation phase. To this purpose, we study four additional distributions computed on the optimal simulated networks – node degrees, path lengths, local clustering coefficients and component sizes – and compare them to the empirical ones. Like in Section 4.2.2, the blue circles in our plots correspond to the mean values and the error bars correspond to the standard deviations of all the quantities we analyze on the 25 realizations of each optimal simulated collaboration network. In many cases, the error bars are not visible, because the values are very narrowly distributed across these 25 realizations.

For the sake of brevity, we do not report the plots corresponding to all of the 12 collaboration networks that we have studied. We only choose one representative sectoral R&D network, pharmaceuticals, reported in Fig. 4.7, and one representative co-authorship network, other applied and interdisciplinary physics (the field containing network theory), reported in Fig. 4.8.

Remarkably, we find that the distributions generated by our model exhibit a good overlap with the corresponding empirical ones, even though the model was not required to reproduce them. Namely, we can retrieve the typical right-skewed degree distribution, the local clustering coefficient distribution, the path length distribution peaked around a mean value of 5 for the pharmaceuticals R&D network and 8 for the co-authorship network in interdisciplinary physics. The model can also reproduce the component size distribution, in particular showing the emergence of a giant component in both networks, together with many smaller components down to size two. The same finding, even though we do not show it here, holds for the remaining collaboration networks we have tested.

Going further in this extended validation procedure, we test the modular properties of the optimal simulated networks, by running the Infomap community detection algorithm on all of their realizations (see Section 4.2). We report the results for the pharmaceuticals R&D network in Fig. 4.9, and for the co-authorship network in other applied and interdisciplinary physics in Fig. 4.10.

Both simulated distributions resemble their empirical counterparts, similarly to the pooled R&D network (see Section 4.2). Another evidence of their similarity is the modularity score

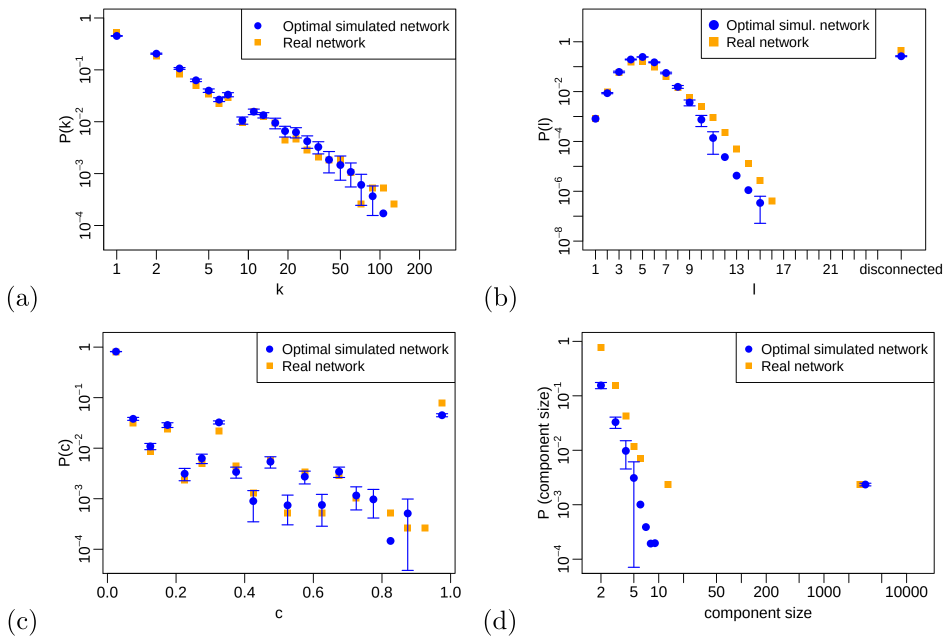


Figure 4.7: Pharmaceuticals R&D network (SIC code 283). Distributions of node degrees (a), path lengths (b), local clustering coefficients (c) and component sizes (d) for the real and the 25 optimal simulated networks.

of the optimal simulated networks – $Q^* = 0.61 \pm 0.01$ for the pharmaceuticals network, and $Q^* = 0.87 \pm 0.01$ for the co-authorship network in interdisciplinary physics. These values are surprisingly close to their empirical equivalents, 0.62 and 0.92 respectively. In all cases, the modularity scores are significantly greater (with a p-value computationally indistinguishable from zero) than the ones obtained for a set of 500 randomly generated networks with the same degree sequence, proving that the modularity is not an artifact of the network size and density; see Chapter 3 for more numerical examples.

In addition, we find that our node labels are actually able to reproduce such a modular structure of both networks. Similarly to the approach described in Section 4.2, we estimate the overlap between the clusters detected via the Infomap algorithm and the circles of influence defined by our node labels, by using the normalized mutual information coefficient I_{norm}. We obtain a striking $I_{\text{norm}}(\text{Labels, Infomap clusters}) = 0.887 \pm 0.003$ for the pharmaceuticals network, and $I_{\text{norm}}(\text{Labels, Infomap clusters}) = 0.952 \pm 0.002$ for the co-authorship network in interdisciplinary physics.

Finally, we test the optimal simulated networks with respect to the distribution of path lengths between every pair of nodes at the moment of the link formation. This should not be confused with the path lengths analyzed before, whose distribution was computed on the final aggregated networks. Now, we only consider pair of nodes that eventually form a link between each other and plot the path length between these nodes at the time step preceding the link formation (see Section 4.2 for more details). We show our findings in

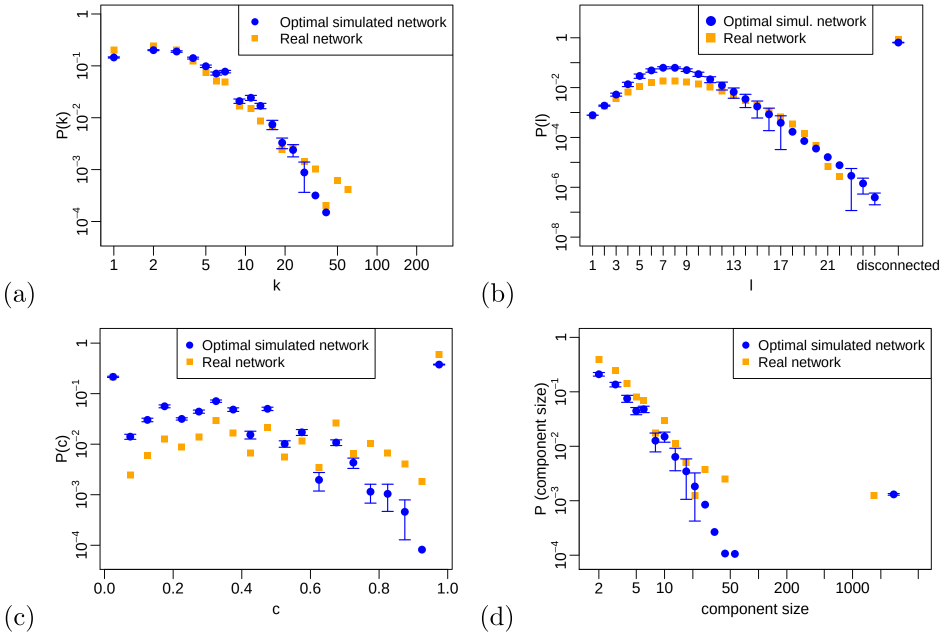


Figure 4.8: Other applied and interdisciplinary physics co-authorship network (PACS number 89). Distributions of node degrees (a), path lengths (b), local clustering coefficients (c) and component sizes (d) for the real and the 25 optimal simulated networks.

Fig. 4.11 for the pharmaceuticals network, and in Fig. 4.12 for the co-authorship network in interdisciplinary physics.

The model can nicely reproduce such microscopic and dynamic feature for both collaboration networks under examination, in particular the counts of links formed between (i) agents belonging to the same connected component of the network, (ii) agents belonging to different disconnected components and (iii) involving at least one newcomer (isolated) agent. Furthermore, the model correctly predicts the formation of links between nodes that are already in the same network component, thus allowing the exact calculation of the shortest path length at the moment of link formation. Both collaboration networks exhibit the tendency of having more links between nodes with a short geodesic distance; in particular, the co-authorship network in other applied and interdisciplinary physics presents many counts for low path lengths, especially 2 (that is, triadic closure) and 1 (that is, repeated collaboration). This feature, which is nicely reproduced by our model, is in agreement with the finding of high probabilities for expert scientists in co-authorship networks to connect with scientists from the same circle of influence – or, in our network representation, high probability for labeled nodes to connect with nodes sharing the same label.

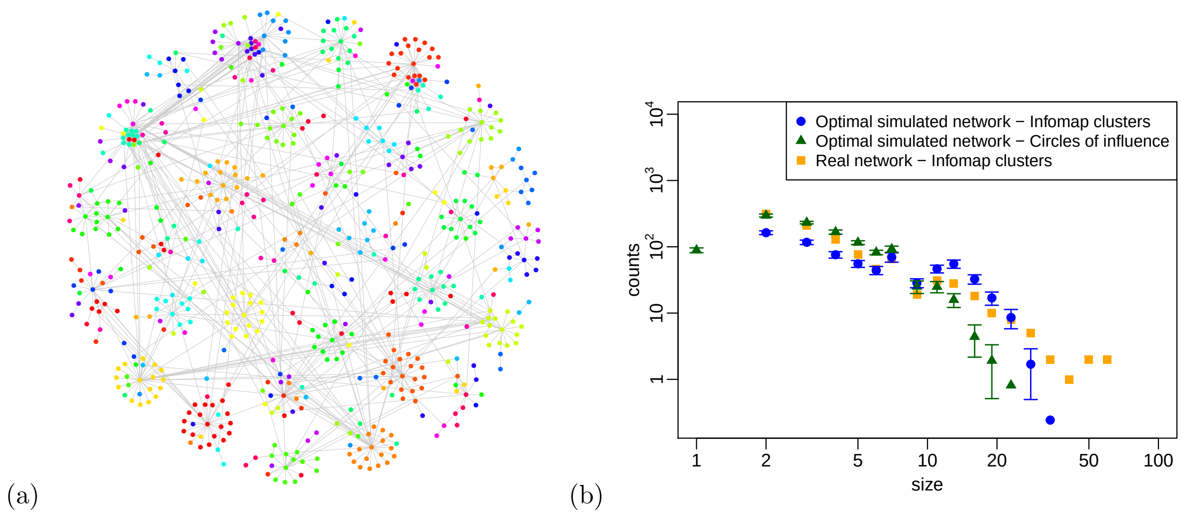


Figure 4.9: Pharmaceuticals R&D network (SIC code 283). (a) Visual representation of one realization of the optimal simulated network including the 30 largest clusters detected by the Infomap algorithm. Distinct clusters are represented by node groups in distinct regions of the plot area. In addition, we depict our node labels by using different colors: most of the nodes in a given cluster share the same label. (b) Size distribution of i. the circles of influence in the 25 realizations of the optimal simulated network, ii. the Infomap clusters in the 25 realizations of the optimal simulated network and iii. the Infomap clusters in the empirical network.

4.4 Discussion

In the present Chapter, we have developed an agent based model of strategic link formation aimed at reproducing the formation of collaboration networks. Inspired by our empirical findings, especially on R&D networks, we have designed a model where the agents, representing real collaborating agents, are endowed with two key attributes: an activity (representing their propensity to engage in new alliances) and a label (representing their membership in a given circle of influence).

Next, we have proposed a simple yet effective set of microscopic rules to reproduce the topology of the observed networks, including both network-endogenous and network-exogenous mechanisms for link formation. Our model is centered around the assumption that the agents have a membership attribute, that we call *label*. Such attribute can be propagated to other agents as a consequence of a collaboration, thus defining the so called *circles of influence* (groups of nodes sharing the same membership attribute). The model includes different link formation probabilities, that depend on both the collaboration initiator's and its partners' membership attributes.

We have first tested our model against the SDC Platinum alliance dataset. By running extensive computer simulations, we have identified the set of linking probabilities that generates the closest network to the empirical pooled R&D network, with respect to average

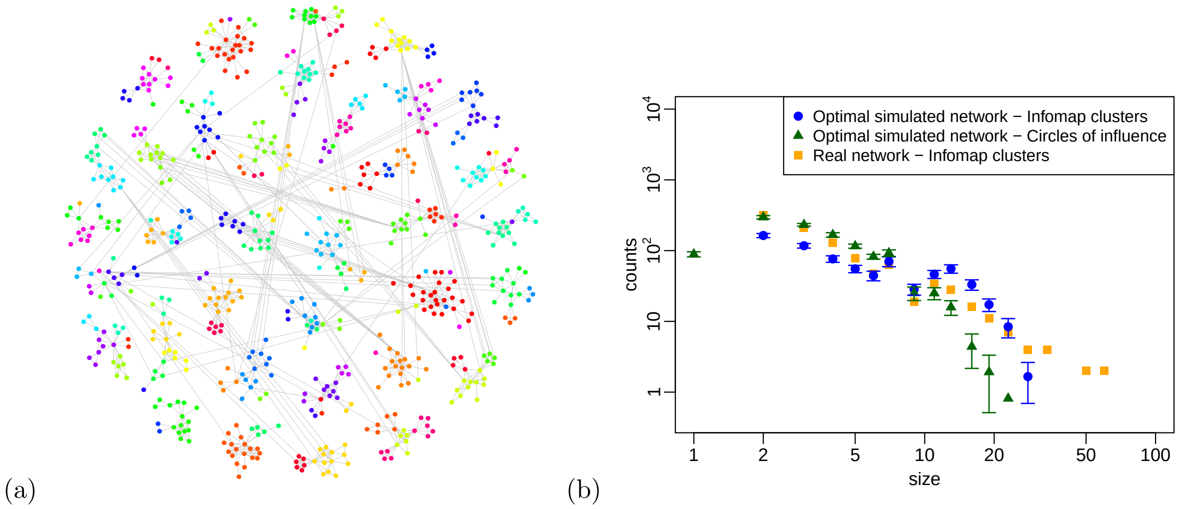


Figure 4.10: Other applied and interdisciplinary physics co-authorship network (PACS number 89). (a) Visual representation of one realization of the optimal simulated network including the 30 largest clusters detected by the Infomap algorithm. Distinct clusters are represented by node groups in distinct regions of the plot area. In addition, we depict our node labels by using different colors: most of the nodes in a given cluster share the same label. (b) Size distribution of i. the circles of influence in the 25 realizations of the optimal simulated network, ii. the Infomap clusters in the 25 realizations of the optimal simulated network and iii. the Infomap clusters in the empirical network.

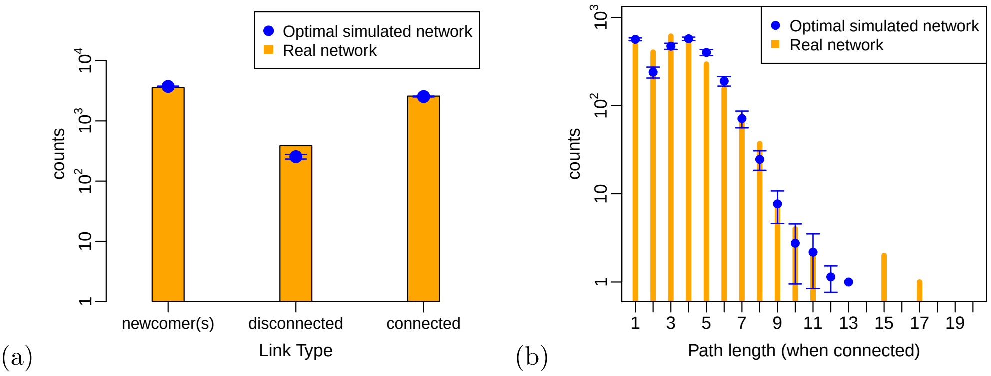


Figure 4.11: Pharmaceuticals R&D network (SIC code 283). (a) Distribution of link types for empirical and simulated networks. “Newcomer(s)” means that at least one of the nodes was isolated (i.e. not yet part of the network) before the link formation; “disconnected” refers to nodes already belonging to the network, but placed in two disconnected components; “connected” refers to nodes already belonging to the same network component prior to the link formation. (b) Distribution of path lengths at the moment of link formation (only for nodes belonging to the same connected component).

degree, global clustering coefficient and average path length. We have found that a labeled node (i.e. an incumbent firm) connects to a node having the same label with probability $p_s^L = 0.3$, to a node having a different label with probability $p_d^L = 0.3$ and, consequently, to

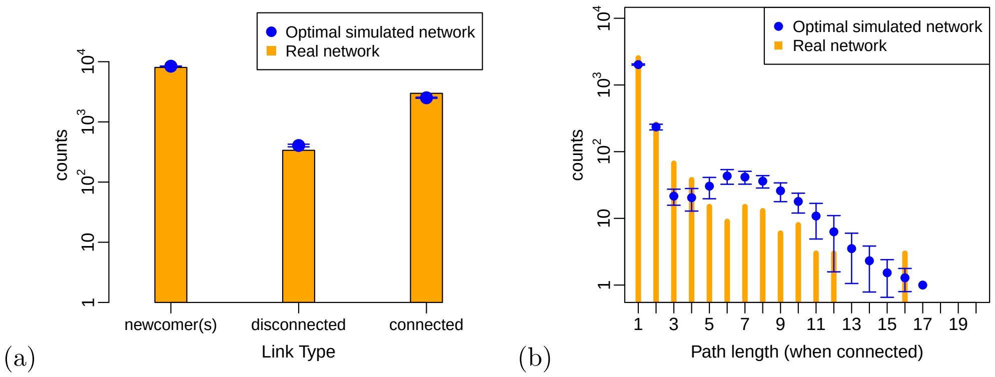


Figure 4.12: Other applied and interdisciplinary physics co-authorship network (PACS number 89). (a) Distribution of link types for empirical and simulated networks. "Newcomer(s)" means that at least one of the nodes was isolated (i.e. not yet part of the network) before the link formation; "disconnected" refers to nodes already belonging to the network, but placed in two disconnected components; "connected" refers to nodes already belonging to the same network component prior to the link formation. (b) Distribution of path lengths at the moment of link formation (only for nodes belonging to the same connected component).

a non-labeled node (i.e. a newcomer firm) with probability $p_n^L = 0.4$. A non-labeled node (a newcomer), when initiating a collaboration, connects instead to a labeled node with probability $p_l^{*NL} = 0.75$ and to another non-labeled node with probability $p_{nl}^{*NL} = 0.25$. The optimal simulated network generated by our model exhibits network measures that deviate from the empirical values by less than 2%.

The linking probabilities listed above have a precise meaning in terms of strategies pursued by firms. Our findings suggest that incumbent firms tend to have a preference towards other incumbent firms: 60% of their alliances belong to this category, split between a 30% probability to connect to a node in the same circle of influence and a 30% probability to connect to a node in a different circle of influence. This finding is in agreement with well-known economic theories (Ahuja, 2000b; Gulati, 1995a) pointing out that previous network connections positively affect the likelihood of alliance formation between two companies. Moreover, we extend previous empirical results (Rosenkopf and Padula, 2008) by including an explicit quantification of the linking probabilities. We find that incumbents willing to form alliances with other incumbents equally share their preferences between firms belonging to the same circle of influence and firms belonging to a different one. In the remaining 40% of the cases, incumbents form alliances with newcomers: these alliances are driven only by exogenous factors (Rosenkopf and Nerkar, 2001), since there cannot be any network endogeneity affecting nodes that are not part of the network yet.

On the other hand, newcomers have a more unbalanced alliance strategy, given that they link to incumbent firms in 75% of the cases. Such alliances are driven by network endogenous factors, namely the newcomers' motivation to join the R&D network by partnering with firms that are already part of it. This is in line with a number of studies (Podolny, 1993; Raub and Weesie, 1990) that have analyzed how being embedded in the network signals attractiveness, also beyond the firm's circle of influence and even to newcomer firms. Indeed, the preferred way for the newcomers to enter the R&D network is to form an alliance with an incumbent firm. Our results confirm and extend previous findings (Powell *et al.*, 2005; Rosenkopf and Padula, 2008) that did never quantify such a preference of newcomers towards incumbents.

However, a fraction (25%) of alliances initiated by newcomers are directed to other newcomers. The reasons behind these alliances are not related to network endogeneity, but rather to exogenous factors such as the firms' commercial or technological capital. Some newcomers prefer to join the R&D network by partnering with other newcomers with no network experience (Baum *et al.*, 2000) – this could be the case, for instance, of small start-up companies in highly technologically dynamic environments – rather than engaging in an alliance with an incumbent firm.

Following this validation procedure, we have extended our tests on a large set of sectoral R&D networks as well as co-authorship networks, by computing for every case the set of optimal linking probabilities. Our findings can be summarized as follows:

- For both R&D and co-authorship networks, labeled nodes (incumbents) tend to form links with other labeled nodes ($p_s^{*L} + p_d^{*L} > 55\%$ in all of the examined collaboration networks).
- When forming a link with another labeled node, the collaboration initiator tends to select a node having the same label, i.e. belonging to the same circle of influence ($p_s^{*L} \geq p_d^{*L}$ in all networks). This tendency is less pronounced in the pooled R&D network and the sectoral R&D networks characterized by high technological dynamism, where incumbents exhibit a balanced alliance strategy, and is instead much stronger in the totality of co-authorship networks, where the circles of influence drive the formation of links between incumbents.
- In all co-authorship networks, plus the sectoral network of R&D, laboratory and testing (again, a highly technologically dynamic sector), non-labeled nodes – i.e. newcomers – tend to form their first links with other non-labeled nodes ($p_{nl}^{*NL} > p_l^{*NL}$). Newcomers tend to enter the network by forming a link with other newcomers.
- For the rest the sectoral R&D networks, instead, non-labeled nodes (newcomers) tend to enter the network by forming a link with labeled nodes, i.e. incumbents ($p_l^{*NL} > p_{nl}^{*NL}$).

Overall, the fine tuning of our model suggests that endogenous mechanisms for network formation are predominant over the exogenous ones, or – in other words – the existing network structures explain most of the newly formed links. This holds for both sectoral R&D networks and co-authorship networks.

However, while newcomers tend to form their first links with other newcomers in co-authorship networks, they instead tend to enter the network by forming links with incumbents in R&D networks. We argue that this is due to higher entry barriers in economic systems than in academic environments. This finding is consistent with empirical evidence; unlike newcomer firms, which join the R&D network for the first time by partnering with incumbent firms, a young scientist writes his/her first paper mostly with other young scientists, being only a small part of the co-authors expert scientists (typically, post-docs or the professor in the same research group).

Finally, we have performed further tests to check whether the model is able to reproduce a set of microscopic network properties, even without imposing any equivalence in the validation procedure. For all examined collaboration networks, we have obtained a surprising agreement with the empirical data. Our model, fed with the optimal parameter combinations, is able to reproduce the distributions of degrees, path lengths, local clustering coefficients and network component sizes. We have also retrieved the distribution of path lengths between every pair of nodes at the moment of link formation, especially including the counts for path lengths 1 (i.e. repeated collaborations) and 2 (i.e. triadic closures). This strongly supports the goodness of our model microscopic rules.

In addition, we have reproduced the emergence of clusters in our collaboration networks. Interestingly, we have found a remarkable overlap between the network partition defined by a widely used community detection algorithm (Infomap) and the one defined by our node labels (i.e. membership attributes). Such overlap, measured through a normalized mutual information criterion, is around 90% for all collaboration networks. We argue that this highlights how effectively the label propagation mechanism can model the formation of agents' circles of influence within every collaboration network. This result is even more remarkable if we consider that the Infomap algorithm detects structural clusters based on the probability flow of random walks in the network (Rosvall and Bergstrom, 2008), while our label propagation mechanism consists of an assignment of a fixed membership attribute – which is not only closer to a real phenomenon, but also computationally easier.

In conclusion, we argue that our model is able to reproduce the formation of links and more complex structures in many R&D and co-authorship networks. The analysis of patterns in the numerical values of the optimal linking probabilities has provided us with insights into the microscopic rules driving the establishment of collaborations in different systems.

The first logical extension to the present model, which we will develop in the next Chapter, consists in a more rigorous definition of the exogeneity rules. This will result in a precise quantification of the technological capital – or knowledge basis – of the agents, and the subsequent effects on link formation and evolution.

Chapter 5

Modeling the exchange of knowledge in a collaboration network

Summary

Following the modeling approach introduced in Chapter 4, we now investigate one mechanism that co-evolves and exhibits complex interdependencies with respect to the network topology, namely the exchange of knowledge in a collaboration network. We include this ingredient in an agent-based model, together with a set of basic network properties, as a first step toward a comprehensive modeling framework. The agent based model we develop here assumes that a knowledge exchange may take place as a consequence of the formation of a link. This allows us to study the complex interdependencies and mutual feedbacks between the network structure and the nodes' intrinsic characteristics (i.e. their knowledge basis). The model parameters that determine the overall properties of the system are the link rewiring rate of the network and the agents' interaction radius. We define the *performance* of the collaboration network as the distance traveled by all of its agents in a knowledge space, that for the sake of simplicity we model as a metric one. We find that, depending on the values of link rewiring and interaction radius, the agents tend to cluster around one or a few attractors in the knowledge space, whose position is an emergent property of the system. And, more importantly, we find that there exist optimal values for both parameters maximizing the network performance.

Based on M. V. Tomasello, C. J. Tessone, F. Schweitzer, "Network dynamics and the exploration of knowledge in collaboration networks", manuscript pending submission. M.V.T. contributred to designing the microscopic rules of the model, made the major effort in running the computer simulations, analyzed all the results and produced all the plots. The manuscript was mostly written by M.V.T.

5.1 Building blocks of the agent-based model

In Chapter 4, we have modeled the formation of collaboration networks inspired by the experimental findings presented in Chapters 2 and 3. The agent based model we have developed, however, considers all the mechanisms of link formation that are not directly attributable to the existing network structure as “exogenous”. The purpose of this Chapter is to understand one of these mechanisms, that has complex interdependencies and co-evolves with the network topology, and include it in an agent based model together with a set of basic network properties, as a first step towards a comprehensive modeling framework. Namely, the aspect we want to investigate now is the knowledge exchange in a collaboration network. In the model that we develop here, such an exchange is assumed to occur as a consequence of the formation of a link; however, only some links actively contribute to this knowledge exchange mechanism, thus introducing complex mutual feedbacks between the network structure and the nodes’ intrinsic characteristics (i.e. their knowledge basis).

The model that we introduce here perfectly suits the description of a system where the collaborating agents have a measurable knowledge basis, such as an R&D network between firms (whose knowledge is well proxied by their patenting activity), but it can be extended – in principle – to any network involving a learning process when links are formed. In a co-authorship network, for instance, scientists learn from each other when co-authoring a paper; however, an empirical validation of the model would be challenging, given that knowledge classifications are typically applied to scientific papers (i.e. the links of the network) rather than the scientists themselves (i.e. the nodes of the network). For this reason, in the continuation of the current chapter, we will often refer to R&D networks as the starting point for the definition of the model’s microscopic rules, laying also the foundations for the empirical validation of the model itself. Likewise, we will use the terms *agents*, *nodes* or *firms* exchangeably. Nevertheless, the reader has to keep in mind that the model can be extended to any collaboration network and even empirically validated, if appropriate data are available.

5.1.1 Model foundations

The model we propose follows an existing stream of literature in the direction of bounded confidence and continuous opinion dynamics models (Axelrod, 1997; Deffuant *et al.*, 2000; DeGroot, 1974; Hegselmann and Krause, 2002; Schweitzer and Behera, 2009), especially applied to innovation networks (Baum *et al.*, 2010; Fagiolo and Dosi, 2003). In the wake of this previous work, we assume that the collaborating nodes are endowed with an evolving knowledge basis, that affects alliances and – in its turn – is affected by them. However, differently from the studies that have been done so far, our model does not focus on the formation of consensus clusters (see Axelrod, 1997; Groeber *et al.*, 2009, in the case of social systems) or technology islands (see Fagiolo and Dosi, 2003, for an economic system). Also, our work differs from previous studies (Gersbach and Schmutzler, 2003; Suzumura, 1992) that are focused on strategic decisions made by firms and the effects that these have on the innovation incentives for the involved firms. We rather focus on the dynamics that leads the system to the observed final state, with emphasis on the "exploration" of the knowledge space by the collaborating agents. We then investigate the existence of an "optimal" network dynamics that maximizes such knowledge space exploration.

With respect to R&D networks, we have shown in Chapter 2 that – despite long-term simultaneous fluctuations – different industrial sectors exhibit different characteristics in their alliance activity (size and density of the corresponding inter-firm network, heterogeneity of degree distributions, other sophisticated topological network properties and so on). Rosenkopf and Schilling (2007) explained part of these observed differences with the so-called "technological regime" of the sector. A technological regime is defined (Nelson and Winter, 1982) as the pattern of behaviors and common practices in an industrial sector, that are influenced by factors such as technological dynamism, technological uncertainty or separability of innovation activities. In the literature, two technological regimes have originally been detected (see Winter, 1984): an *entrepreneurial regime*, where R&D activities are mainly carried out by new innovative firms, and a *routinized regime*, where innovation is mainly done by incumbent firms. These two extremes are often referred to as *explicit knowledge regime* and *tacit knowledge regime*, respectively, because firms in the network tend to interact with diverse firms or with similar firms (in terms of knowledge basis), in the respective cases. However, this distinction has been extended over the years, bringing to the identification of several classes of technological regimes, spanning between the two aforementioned extremes.

Therefore, the model we present is intended to reproduce the knowledge exchange process occurring in a collaboration network. Our aim is to capture the existence of an optimal rate of alliance formation and its dependence on the underlying technological regime.

5.1.2 Microscopic rules: stylized facts and theoretical arguments

The microscopic rules of our model are inspired by a number of stylized facts, as well as theoretical speculations, in network evolution studies, opinion dynamics models, R&D and collaboration networks. Below, we provide a brief description of every building block that we employ in the definition of our agent-based model.

Monogamous network approximation. Inter-organizational networks are proven to have low density, i.e. only a small fraction of all potential collaborations between companies are actually realized. The density of R&D networks ranges from 0.1% to 1% for all industrial sectors, as we have shown in Chapter 2. This empirical evidence allows us to model the formation of R&D alliances between companies as a *monogamous network*, i.e. a network in which every agent is linked to only one other agent at every time step (Vazquez and Zanette, 2010). Furthermore, the degree distributions (see Chapters 2 and 3) show that the vast majority of the nodes have only one partner, even in the cumulative representation. Even though all agents can have only one link at every time step, they are allowed to change their partners in the subsequent time steps and, depending on their position, they can actually collaborate with many firms in a small time window. To have a more realistic picture, it would be possible to aggregate an appropriate number of network snapshots over time, as suggested by Baum *et al.* (2010) to study the topological properties of theoretical R&D networks; however, this investigation is beyond the scope of the present model. This assumption might seem strong for high-technology industries, such as Pharmaceuticals or Computers, that – although having low density – show small world properties and hierarchical structures. The "hubs" of these industries can actually have more than a hundred partners at the same time, with which they collaborate on different projects (Hanaki *et al.*, 2010; Powell *et al.*, 2005). However, we argue that the monogamous assumption holds, because we are still able to capture these cases of enormous alliance activity in our model through the rewiring of the links.

Position of companies. In the knowledge-based view of the firm, every company is endowed with a knowledge basis that uniquely identifies its resources and its capabilities. We assume that a firm is represented by an agent in our modeling framework, and associate it with a vector of D components, each of which represents its level of knowledge in a given direction. Furthermore, we directly associate these vectors to a metric *knowledge space* in which the collaborations occur: every firm occupies a point in this $D-$dimensional space, whose coordinates are given by its knowledge vector. Such an approach is similar to a more general model (in the broader context of social influence), that is the one proposed by Axelrod (1997). The concept of a metric knowledge space has been used by Groeber *et al.* (2009), in one dimension, and by Baum *et al.* (2010); Fagiolo and Dosi (2003), in two dimensions. We generalize the dimensionality of the space to D.

Alliance formation. In our monogamous network, all nodes are linked in pairs at every time step. We assume that two pairs of allied nodes mutually rewire their links at every time step with a given probability, and the new formed links are active if the Euclidean distance between the new partners is *smaller than a threshold value*. Such a proximity condition models the theoretical argument in Cohen and Levinthal (1989) and Cohen and Levinthal (1990), highlighting that an interaction between two companies is profitable only if their *absorptive capacity* is large enough or – in other words – their knowledge distance is small enough. The choice of the Euclidean metric to compute this distance is quite realistic, even if it implies extensive information of the companies about their mutual position in the knowledge space. Indeed, obtaining detailed information about a company, its patent production, its scientific production and its activities in general is nowadays not only feasible – thanks to the Internet – but actually done by most of the firms willing to engage in an alliance (see Ahuja, 2000a; Baum *et al.*, 2010; Sampson, 2007). The threshold value is instead supposed to model the technological regime characterizing the collaboration network under examination. The larger its value (corresponding to a more explicit knowledge regime), the more easily the agents establish active collaborations, even if their knowledge distance is large. The smaller its value (corresponding to a more tacit knowledge regime), the closer the agents have to be in the knowledge space in order to establish an active collaboration.

Partner selection. The dynamics of alliance formation in the present model is assumed to be *semi-random*, meaning that the rewiring of links between nodes occurs randomly and independently of the position of the nodes themselves in the knowledge space: we call this an *exploration phase*. However, a link between two nodes is *active* only if they are close enough in the knowledge space: if this happens, a so-called *knowledge transfer phase* begins. The rewiring mechanism does not want to be a close representation of what happens in reality. It only has the function of modeling the volatility of R&D alliances, capturing the characteristic time scale at which firms decide to engage in a new alliance. The second focal aspect that we want to model – namely the formation of alliances at the right knowledge distance – is instead fully captured by the threshold value for the knowledge distance of the potential partner.

Approaching in the knowledge space. Once a link has been established, we assume that a knowledge exchange between the partners takes place, causing their knowledge bases to become more similar and bringing them closer in the knowledge space. This assumption is in line with the conceptualization of R&D alliances as a means to exchange technological knowledge among firms (Gomes-Casseres *et al.*, 2006; Grant and Baden-Fuller, 2004; Mowery *et al.*, 1998; Owen-Smith and Powell, 2004) and has already been used in a number of agent based models (Cowan *et al.*, 2007; Gilbert, 2004; Pyka and Fagiolo, 2007). The speed at which the agents approach each other in our agent-based model represents another parameter of the network dynamics. Our work studies a scenario in which the two partners approach with respect to all D dimensions of the knowledge space (as done by Baum *et al.*, 2010). That is, we assume that knowledge spillovers occurring in a R&D alliance cause the partners to exchange knowledge along every dimension, not limiting the knowledge transfer to a specific R&D project that they have in common.

Exploration of knowledge space. Finally, we want to study the performance of the whole collaboration network as a function of the relevant model parameters. The indicator we propose to measure the global knowledge production depends on the exploration of the knowledge space by the collaborating agents: in other words, it quantifies the distance traveled by all agents during the whole simulation. In our model, we consider that knowledge itself is represented by the motion in the space, which is fully captured by this indicator. The underlying assumption is that the exploration of as many locations as possible is beneficial for the collaboration network, in that it allows the agents to come in contact with many technological opportunities, potentially leading to more frequent innovations (Fagiolo and Dosi, 2003). Testing our model by means of computer simulations, we find that the rewiring of links and the mutual knowledge exchanges over time eventually lead the whole system to a steady state through a non-trivial dynamics. The model and its results are presented in detail in the next two Sections.

5.2 Development of the agent-based model

Starting from the evidence and the arguments presented in the previous Section, we now present the implementation of the agent-based model. We consider a network composed of N nodes, each representing an agent – in the particular case of R&D networks, a firm – performing collaboration activities in a knowledge space. The model is implemented by means of computer simulations, consisting of a sequence of discrete time steps of length $\mathrm{d}t$. The microscopic interaction rules are described below.

5.2.1 Exploration phase

Every node i is located in a metric space (henceforth, the knowledge space); this point has coordinates $\mathbf{x}_i$, identified by a vector of D real numbers ranging from 0 to 1. The coordinates of every node can be thought of as the ratios of the corresponding firm's expertise along each of the D dimensions of the knowledge space. At the initial stage of every simulation, all the nodes' positions are drawn from a uniform distribution.

$$\mathbf{x}_i \equiv (x_{i1}, x_{i2}, \ldots, x_{iD}) \qquad i = 1, \ldots, N. \tag{5.1}$$

All nodes in our R&D network have the possibility to change their partner, thus generating a dynamic network topology. We model this by means of a link rewiring mechanism. The time steps in our computer simulations have a duration equal to $\mathrm{d}t$; in each time step, two pairs of connected firms are randomly chosen and, with a rate λ, they rewire their links. We call this process "exploration phase", and depict it in Fig. 5.1. Let us assume that the nodes i and j and the nodes i' and j' constitute the two linked pairs chosen at time t.

With probability $\lambda \mathrm{d}t$, they mutually exchange their partners, and at time $t + \mathrm{d}t$ the nodes i and i' and the nodes j and j' will form the new linked pairs. With probability $1 - \lambda \mathrm{d}t$, instead, nothing happens and at time $t + \mathrm{d}t$ the nodes i and j and the nodes i' and j' will still be respectively linked.

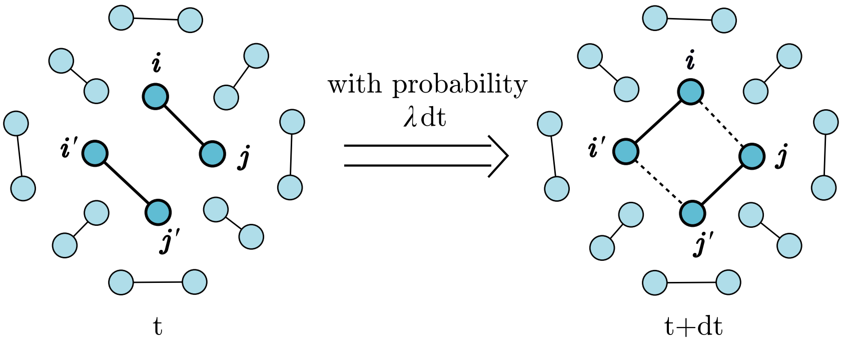


Figure 5.1: Schematization of a link rewiring between two pairs of connected nodes. At time t, the nodes i and j and the nodes i' and j' are linked in pairs. These two couples of nodes are selected and, with probability $\lambda \mathrm{d}t$, they switch links: at time $t + \mathrm{d}t$ the nodes i and i' and the nodes j and j' are the new linked pairs. Obviously, with probability $1 - \lambda \mathrm{d}t$, no rewiring happens.

Such a random search for partners in the exploration phase might seem to be an unrealistic assumption; however, this has the only function to model the volatility of R&D alliances, or collaborations in general, capturing the characteristic time scale at which an agent decides to engage in a new collaboration. The rate λ can be indeed thought of as the inverse of the characteristic time elapsed before a firm takes part in a new alliance. Even though the potential partner is selected at random, the R&D alliance – or the collaboration – will be actually "active" only if the partner fulfills a certain proximity condition in the knowledge space, as we will explain below. Therefore, such exploration is not fully arbitrary, and leads to the establishment of an actual collaboration only under specific conditions. It is worth mentioning that the results of our simulations remain qualitatively unchanged if we use any different random link creation process, or if we relax the monogamous network assumption (compatibly with Tessone and Zanette, 2012).

5.2.2 Knowledge transfer phase

The whole linking and rewiring process in our model occurs independently of the node knowledge positions, but their distance in the knowledge space has a determinant effect on the subsequent network dynamics. Indeed, the key ingredient to our model is the existence of an optimal absorptive capacity for a profitable R&D alliance between two firms. We assume that a link is *active* if the corresponding node pair exhibits a knowledge distance smaller than a given threshold value. If this proximity condition is not fulfilled, even though the corresponding nodes are connected, their link is considered to be *inactive*, causing no effect at all on the system. The proximity condition is evaluated for every pair of linked nodes i and i' as follows:

$$|\mathbf{x}_i(t) - \mathbf{x}_{i'}(t)| < \varepsilon\sqrt{D} \tag{5.2}$$

where we employ the Euclidean distance $|\cdot|$, consistently with the assumption of evaluating the diversity of each firm's knowledge portfolio in all dimensions. $\sqrt{D}$ is the maximum possible distance between two points in a D-dimensional Euclidean space. The parameter ε, ranging from 0 to 1, is the threshold interaction radius inside which nodes are able to interact and collaborate profitably. Only links whose corresponding nodes fulfill this proximity condition are considered to be *active*. Such an interaction radius can be associated with the knowledge regime characterizing the collaboration network under examination. A large ε means that the firms can potentially see and explore a large portion of the knowledge space, being the knowledge highly *codified*. A small ε represents instead a regime of *tacit* knowledge, where firms are able to establish alliances only if their technological positions are already close.

We assume that an R&D alliance causes the two involved firms to pool their resources and their knowledge basis, thus approach along every dimension in the knowledge space. Thanks to *knowledge spillovers*, both firms will acquire common practices or a shared jargon, not limiting the knowledge transfer to that specific R&D project that they have in common, as previously discussed.[1] If i is an agent and i' is its unique partner in the collaboration network at time t, both will move towards each other by identical paths in the knowledge space, provided that the proximity condition expressed in Eq. 5.2 holds. The model dynamics equation is the following:

$$\dot{\mathbf{x}}_i(t) = \mu\,[\mathbf{x}_{i'}(t) - \mathbf{x}_i(t)], \qquad \text{if} \quad |\mathbf{x}_{i'}(t) - \mathbf{x}_i(t)| < \varepsilon\sqrt{D} \tag{5.3}$$

where μ is defined as the *learning rate* of the agents. This parameter is constant over time and for all nodes in the collaboration network, and can be thought of as the propensity of the agents to exchange knowledge with their partners, thus making their knowledge bases more similar over time. It should be noted that the parameter μ is a rate, not a speed; the actual speed at which the corresponding nodes move in the knowledge space is given by the product of the rate μ and their distance: therefore, the farther they are in the knowledge space, the faster they approach. When their distance decreases, so does the potential for new learning from the collaboration, and the approaching speed drops consequently. This

[1] However, we have also tested a scenario in which two allied firms exchange knowledge only in one dimension, thus moving in only one dimension of the knowledge space as well. The results remain qualitatively unchanged.

interpretation is clear in Eq. 5.4, which represents the way we implement the model in computer simulations with discrete time steps of length $\mathrm{d}t$. The evolution of every agent's position $\mathbf{x}_i$ can be expressed as:

$$\mathbf{x}_i(t + \mathrm{d}t) = \mathbf{x}_i(t) + \mu \,\mathrm{d}t\, [\mathbf{x}_{i'}(t) - \mathbf{x}_i(t)] \tag{5.4}$$

We depict such knowledge exchange mechanism in Fig. 5.2. The nomenclature and the meaning of all the model parameters we introduced in this Section are summarized in Table 5.1.

Parameter	Meaning	Type of parameter
N	Number of agents (system size)	Static
D	Dimensionality of the metric knowledge space	Static
ε	Agents' interaction radius (knowledge regime)	Static
λ	Link rewiring rate	Network dynamics
μ	Approaching rate in the knowledge space	Network dynamics

Table 5.1: Model parameters and their description. The "static" parameters are associated with the system structural features, while the "network dynamics" parameters define the characteristic speed at which the system evolves.

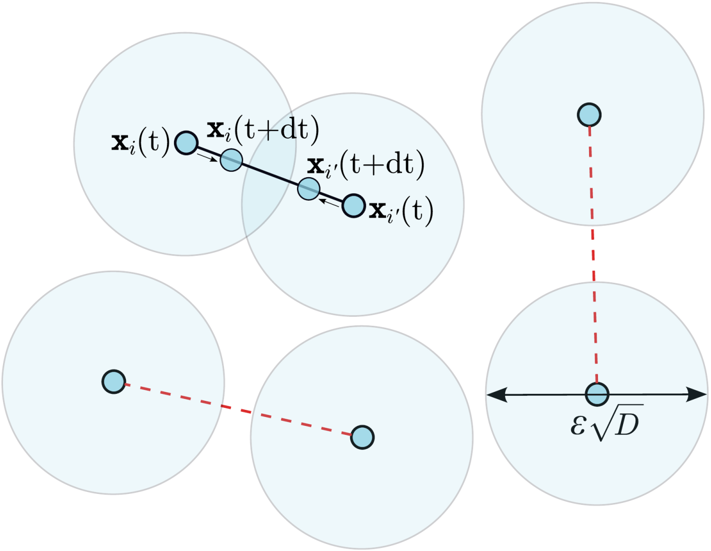


Figure 5.2: Schematization of the knowledge exchange process in a bi-dimensional space ($D = 2$). At time t, the agents i and i' are linked and their distance $|\mathbf{x}_{i'}(t) - \mathbf{x}_i(t)|$ is smaller than $\varepsilon\sqrt{D}$; consequently, at time $t + \mathrm{d}t$, their positions $\mathbf{x}_i(t + \mathrm{d}t)$ and $\mathbf{x}_{i'}(t + \mathrm{d}t)$ will approach in the knowledge space. The picture includes other pairs of connected agents, whose distance is larger than $\varepsilon\sqrt{D}$. Therefore, these links are *inactive* (depicted in dashed red lines) and do not originate any motion in the knowledge space.

5.3 Results

We have performed extensive computer simulations, by applying the dynamics presented in Sec. 5.2 and varying the values of the relevant model parameters. In particular, we vary the size N of the network from 10 to 2000 nodes, the dimensionality D of the knowledge space from 1 to 50, the interaction threshold radius ε from 0 to 1, the learning rate μ from 10^{-3} to 10^{3} and the rewiring rate λ from 10^{-5} to 10^{5}. All of these parameters are explored in discrete intervals, whose width is appropriately chosen – as we discuss below in more detail.

Main model parameters and their meaning. We argue that the network evolution is essentially characterized by two driving forces with overall opposite effects. The first one is the formation of active links (i.e. the establishment of profitable alliances or collaborations); this force tends to push agents closer in the knowledge space, given the resulting approaching motion. The second force is the link rewiring (representing the dissolution of old collaborations and the formation of new ones), that stimulates agents to explore new portions of the knowledge space. This force could result in an faster overlap of every agent's knowledge position, but it could also result – under certain conditions – in preventing the agents from converging to a knowledge attractor, thus keeping them far-between in the knowledge space.

These competing forces are associated with the two model dynamics parameters, respectively the approaching rate μ and the link rewiring rate λ. However, it is clear that the relation between these two parameters will substantially affect the emergent properties of the system. What truly affects the resulting dynamics of the network are not the absolute values of the two rates μ and λ, but the ratio of the two. Indeed, using a configuration with the same μ to λ ratio, but with smaller absolute values, will only lead to a longer computer simulation (i.e. more discrete time steps are needed), without qualitatively changing the results. Therefore, in the continuation of the current chapter we present our findings by keeping the value of of the learning rate fixed to $\mu = 1$, and studying the effect of the dynamics parameter λ only.

The second relevant model parameter on which we focus our attention is the threshold interaction radius ε, a static parameter representing the knowledge regime in which the collaborating agents move. We explore a series of values ranging from a totally tacit knowledge regime ($\varepsilon = 0$) to a totally explicit one ($\varepsilon = 1$).

Network performance. The variable that we investigate as indicator of the network performance is the *mean knowledge path* $\langle K \rangle$ of the collaborating agents. We define the path covered by every agent in the knowledge space K_i as the sum of all the distances that the agent travels in every time step of the simulation:

$$K_i = \int_{t=0}^{T_{\max}} |\dot{\mathbf{x}}_i(t)| \, \mathrm{d}t \tag{5.5}$$

where $T_{\max}$ is the duration of an entire computer simulation. It should be noted that the measure $|\dot{\mathbf{x}}_i(t)| \, \mathrm{d}t$ is a positive scalar and expresses the actual distance traveled by the agent i, differently from its net displacement $\dot{\mathbf{x}}_i(t) \, \mathrm{d}t$, which is a vectorial quantity. The measure K_i is then averaged over all the N network agents to obtain the mean knowledge path $\langle K \rangle = N^{-1} \cdot \sum_i K_i$. We hypothesize that this measure can provide a meaningful indication of the system performance, because – as already discussed in Section 5.1 – firms are proven to innovate more when they come in contact with more technological opportunities (i.e. they explore the knowledge space). We argue that the same reasoning can be as well extended to other types of collaborations that involve learning and/or knowledge exchange processes.

We present the results in Fig. 5.3, for a representative network of $N = 200$ agents moving in a knowledge space with $D = 10$ dimensions. As already mentioned, the parameter μ is fixed to 1, and we study the dependence of $\langle K \rangle$ on the dynamics parameter λ and the static parameter ε. For a two-dimensional representation of the same results, see Appendix D.

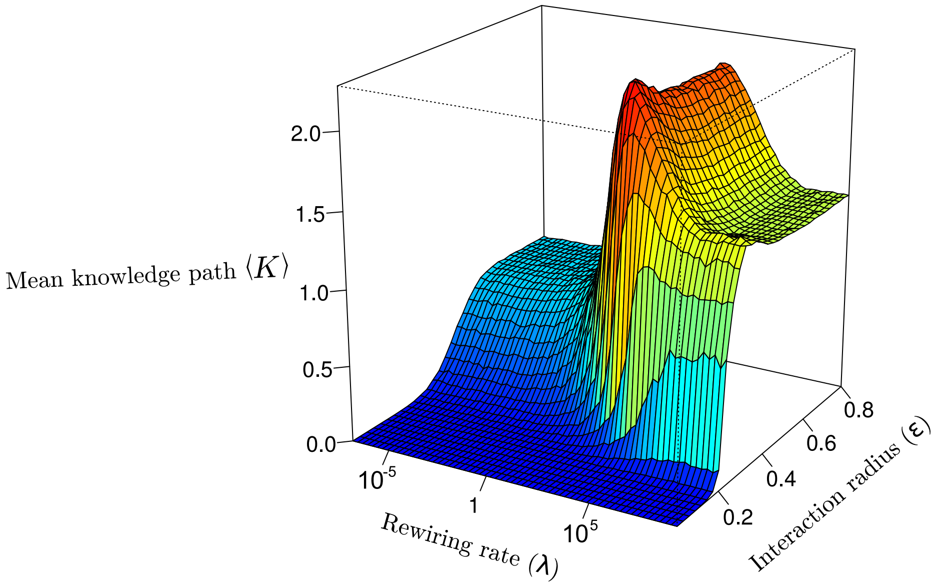


Figure 5.3: Mean knowledge path $\langle K \rangle$ (displayed by means of both the z-elevation and the color scale), as a function of the rewiring rate λ and the interaction radius ε. The R&D network under examination has $N = 200$ nodes and learning rate $\mu = 1$, in a 10−dimensional knowledge space. We generate 1000 simulations for each parameter set and then average the results.

We find that the mean knowledge path $\langle K\rangle$ exhibits a peak in correspondence of two optimal values for both the rewiring rate λ and the interaction radius ε (for the specific case we present in Fig. 5.3, these values are $\lambda = 10^3$ and $\varepsilon = 0.25$). If we take a closer look at the network performance, we find that $\langle K\rangle$ shows a monotonic growing trend as a function of λ, when the interaction radius ε is lower than a certain value ε^* (in our example, $\varepsilon^* < 0.2$). When fixing the interaction radius to larger values $\varepsilon \geq \varepsilon^*$, we do instead find that $\langle K\rangle$ exhibits a non-trivial peak as a function of λ. This means that, as the knowledge regime becomes more explicit, and the agents are allowed to form active collaborations with more diverse partners in terms of knowledge basis, there exists an optimal rewiring rate maximizing the distance actually explored by the agents in the knowledge space.

The behavior of the mean knowledge path $\langle K\rangle$ can be also interpreted as a function of the interaction radius ε, while keeping the rewiring rate λ fixed. What we find is that $\langle K\rangle$ grows with ε to a saturation level (when $\varepsilon > 0.6$), if the rewiring rate is small ($\lambda < 1$, for the case under study). If we fix the rewiring rate λ to a value larger than 1, we find instead that $\langle K\rangle$ increases to a peak, in correspondence to $\varepsilon = 0.4$, and then decreases again to stabilize for $\varepsilon > 0.6$. This means that, when the characteristic alliance rewiring rate of the network is bigger than the characteristic learning rate of the agents, there exists an optimal threshold interaction radius (corresponding to a moderately explicit knowledge regime) maximizing the distance covered by the agents in the knowledge space.

Knowledge clusters. We investigate a second emerging property of the system, namely the *number of knowledge clusters* appearing in the network at the end of every model run. We define a knowledge cluster as a group of nodes whose mutual distances are smaller than ε. Moreover, the distance between every node in that cluster and every node outside that cluster has to be larger than ε, meaning that all the agents in the cluster will asymptotically converge to one attractor and no further inclusion of any other agent in the cluster is possible.

It is clear that the maximum possible value of knowledge clusters equals the number of nodes N; we expect to observe such a value in correspondence with a low value of the interaction radius ε, when the agents are virtually unable to establish active links. Likewise, the minimum possible number of knowledge equals 1; we expect to observe such a value in correspondence with high values for the interaction radius ε, when most established collaborations are active, thus facilitating the convergence of all agents toward one knowledge attractor.

Similarly to the mean knowledge path, we present our results in Fig. 5.4, for a network of $N = 200$ agents in a knowledge space with $D = 10$ dimensions; μ is fixed to 1.

We find that the number of clusters generally increases by decreasing the interaction radius ε. As expected, one extreme case occurs for $\varepsilon = 0$ (completely tacit knowledge regime,

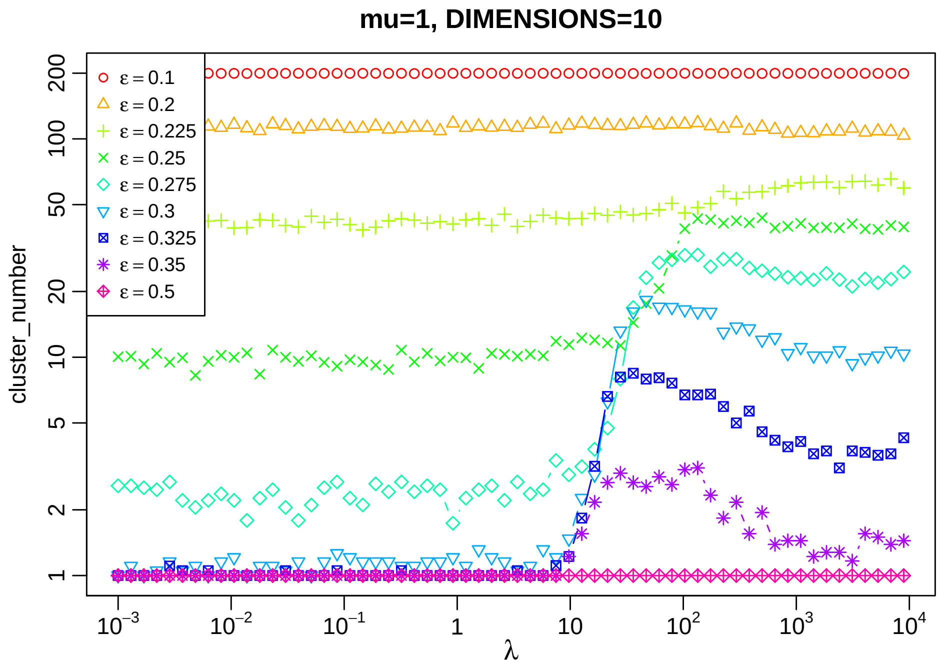


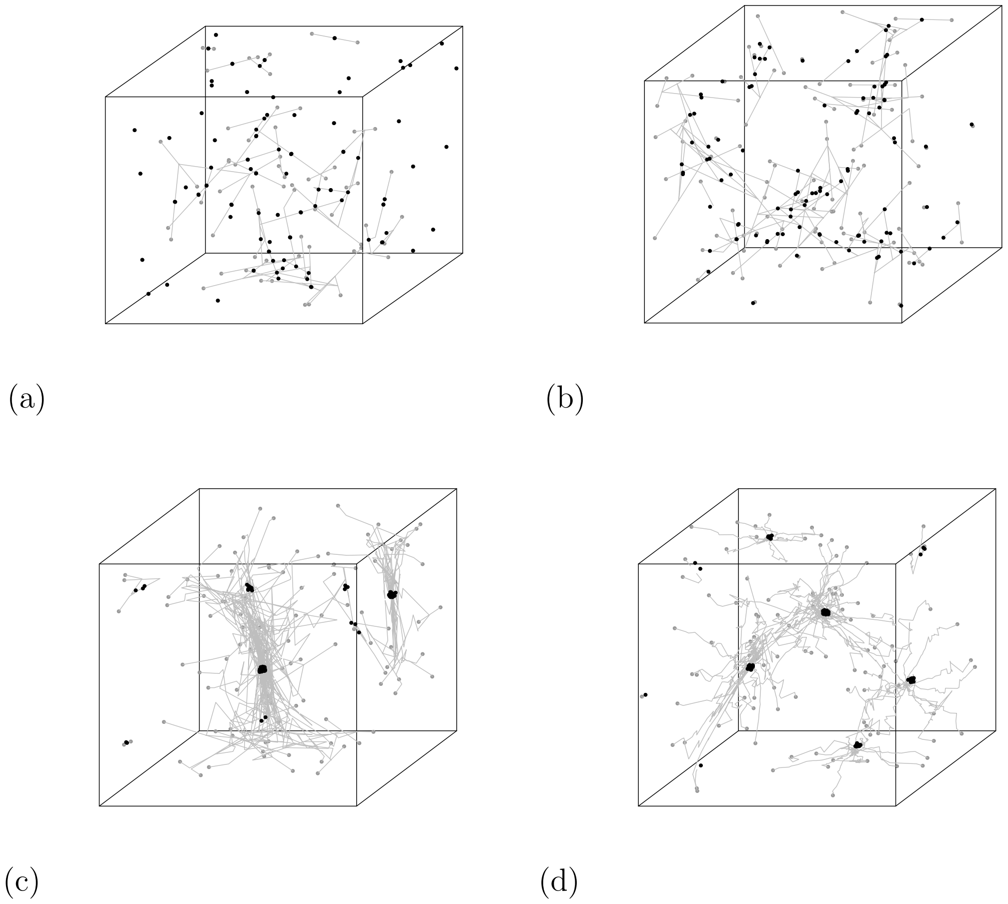
Figure 5.4: Number of knowledge clusters as a function of the rewiring rate λ, for a set of representative values of the interaction radius ε. The network under examination has $N = 200$ nodes and learning rate $\mu = 1$, in a 10−dimensional knowledge space. We generate 1000 simulations for each parameter set and then average the results.

where any interaction is by definition impossible), in which we have as many clusters as agents – independently of the rewiring rate λ. The other extreme case occurs for $\varepsilon \geq 0.5$ (highly explicit knowledge regime), in which all the nodes interact between each other converging in only one cluster – again, independently of λ.

For intermediate values of ε, we observe an interesting dependence of the knowledge cluster number on the rewiring rate λ. When λ is low, we find the existence of one or very few knowledge clusters, because the overall effect of such a slow rewiring rate is that all nodes tend to get closer in the knowledge space before the corresponding links are cut and rewired. As a result, all nodes are eventually part of the same knowledge cluster. From the visual examples in Fig. 5.5 (a) and (b), we can observe that such clusters are dispersed in the knowledge space, and the presence of a central attractor is not visually detectable, even though all the agents are in principle within interaction distance. What happens, in fact, is that every pair of agents converges to the midpoint of the segment connecting them; the system then freezes in this configuration, being the rewiring rate too low to allow for new collaborations and new explorations.

When the value of λ increases, we instead observe a higher number of knowledge clusters. These cluster are well delimited in the knowledge space and, as we show in the examples of Fig. 5.5 (c) and (d), the presence of attractors is visually evident. Such non-trivial effect derives from the fact that the nodes cut their links and form new ones before the approaching mechanism is complete, thus exploring a bigger portion of the knowledge space and ending up in more than one attractor, occupying different regions of the space.

Interestingly, the effect of experiencing more alliances with different partners is therefore the emergence of distinct knowledge attractors, rather than causing all firms to converge towards the same knowledge attractor, thus uniforming their knowledge bases. We report a visual example of cluster emergence as a function of the rewiring rate λ in Fig. 5.5.



(a)

(b)

(c)

(d)

Figure 5.5: Knowledge trajectories for a network with $N = 200$ nodes and learning rate $\mu = 1$. For the sake of visualization, here we use a knowledge space with $D = 3$ dimensions, easily representable as a cube. The initial positions of the nodes are depicted with gray dots, their trajectories with gray lines, and their final positions with black dots. We keep the threshold interaction radius constant to $\varepsilon = 0.3$, and show four cases corresponding to rewiring rate λ equal to: (a) 10^{-4}, (b) 10^{-2}, (c) 10^{-1}, (d) 1.

Convergence time. We find that the networks generated by the model eventually converge to a steady state, in which all the agents occupy one or more fixed positions, and no further collaborations, nor motion in the knowledge space, are possible. In other words, such steady state represents a configuration in which the collaborating agents have depleted all the potential for new knowledge exchange.

We define a convergence criterion based on the agents' motion in the knowledge space, and assume that the steady state is reached if the total knowledge path traveled by all the agents in the last time step is smaller than the 0.5% of the cumulated covered knowledge path. Indeed, all of the network measures described above are computed only after the steady state is reached. We show in Fig. 5.6 the trend of the convergence time as a function of λ and ε, for the same representative network we have studied before.

On the one hand, we find that all the relevant parameter configurations reach a steady state before the computer simulation ends. Indeed, it should be noted that the parameter combinations that are not able to reach a steady state before the end of the simulation (those with $\varepsilon < 0.15$ or generally low λ) are the ones generating the lowest values of mean knowledge path, for the reasons we previously discussed. Therefore, we forcedly stop all computer simulations after $20,000$ time steps, affecting only a small fraction of the parameter space and not influencing our results.

On the other hand, we find an unexpected trend in the convergence time as a function of λ for some parameters combinations. One would expect that the convergence time decreases proportionally to $1/\lambda$, being this quantity (the inverse of the rewiring rate) a measure of the characteristic time of the system for a complete interaction between all agents. However, we observe this trend only for the extreme cases of highly explicit knowledge regimes (where a complete interaction between all agents in the space can take place), corresponding to $\varepsilon \geq 0.5$. What we find is instead a non-trivial trend of the convergence time as a function of λ for all the other values of ε, showing plateaux for high values of λ. This means that the complex network dynamics, in the presence of certain approaching and link rewiring rates, can lead the system to a later convergence than the one suggested by the characteristic rate $1/\lambda$ alone.

5.4 Discussion

In this chapter we have developed an agent based model of dynamic collaboration formation and knowledge exchange. The novel contribution of the model is that it incorporates a process of knowledge exchange and studies its co-evolution and interdependencies with respect to the collaboration network structure.

Studying the interactions of a set of agents in a metric knowledge space, by means of computer simulations, we have found that the system follows a non-trivial dynamics and reaches a steady state in which the agents cluster around a set of emerging attractors. The model parameters that determine the overall properties of the system are the link

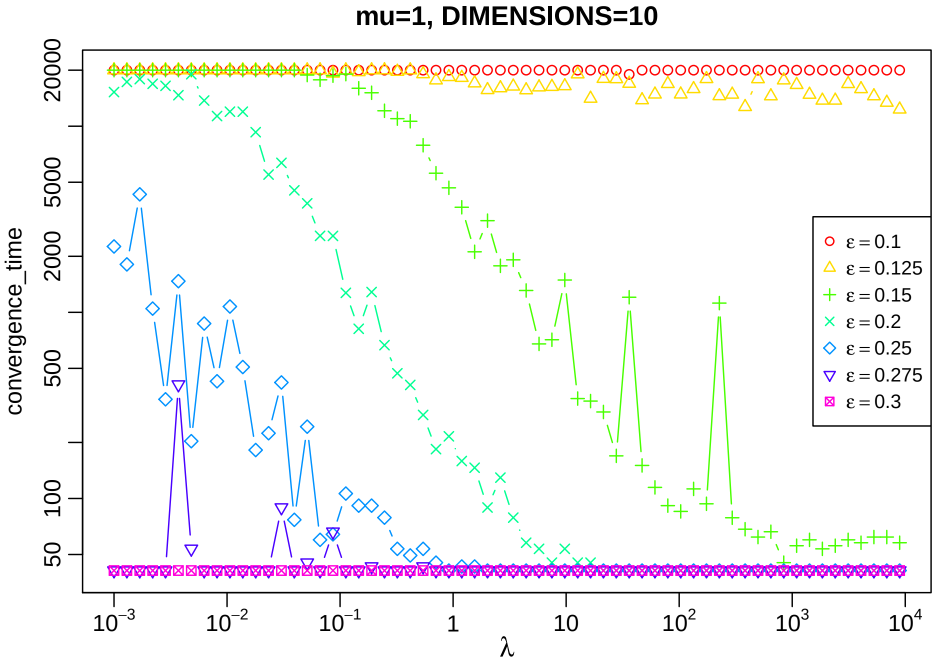


Figure 5.6: Convergence time as a function of the rewiring rate λ, for a set of representative values of the interaction radius ε. The network under examination has $N = 200$ nodes and learning rate $\mu = 1$, in a 10−dimensional knowledge space. We generate 1000 simulations for each parameter set and then average the results.

rewiring rate of the network λ and the agents' interaction radius ε.

We define a knowledge cluster as a group of nodes whose mutual distances are smaller than the threshold interaction radius ε, and whose distance with every node outside the cluster is larger than ε (meaning that all the agents in the cluster will asymptotically converge to one attractor and no further inclusion of any other agent in the cluster is possible). We have found that the number of knowledge clusters observed at the end of the network evolution decreases by increasing the threshold interaction radius ε, because the agents are able to collaborate with partners located farther away in the knowledge space, thus converging all together towards one position. When the knowledge regime is strongly tacit or strongly explicit, the number of knowledge clusters depends only on ε itself, and not on the alliance rewiring rate λ. The most interesting case occurs for intermediate knowledge regimes, in which the number of knowledge clusters increases with λ. Small rewiring rates lead to the emergence of only one knowledge cluster, which is dispersed in the knowledge space and does not clearly exhibit the presence of a knowledge attractor in it. Faster alliance rewiring rates lead the agents to potentially have collaborations with more partners, allowing the emergence of a larger number of knowledge clusters; in this case, the presence of knowledge attractors, around which the firms eventually cluster, is (even visually) clear.

In this model, our underlying assumption is that the exploration of as many locations as possible is beneficial for the entire collaboration network. For this reason, we consider the distance explored by the agents in the knowledge space $\langle K \rangle$ as a performance indicator of the network evolution. We have found that there exists a specific parameter combination maximizing such indicator of performance, specifically intermediate values of both the rewiring rate λ and the interaction radius ε.

In particular, if we focus on the dependence of $\langle K \rangle$ on λ, given a fixed ε, we find that there exists an optimal value λ^* maximizing $\langle K \rangle$. Such optimal rewiring rate λ^* exhibits a weak dependence on ε; namely, it slightly decreases when ε increases (only for intermediate values $0.2 \leq \varepsilon \leq 0.4$). This is consistent with some empirical studies (Gulati *et al.*, 2012; Rosenkopf and Schilling, 2007, e.g.), that show a varying alliance formation rate across industrial sectors. Similarly, we have found that, given a fixed alliance rewiring rate, there exists an optimal interaction radius ε^* maximizing the mean knowledge path $\langle K \rangle$.

While we believe that the study of the knowledge clusters is fascinating, and that the model we have developed could certainly represent a contribution in this direction, we do not further investigate this aspect in the continuation of the present dissertation. We will rather investigate how the collaboration network evolution affects the performance of the whole system. In this respect, our finding of the existence of optimal parameter configurations maximizing the system knowledge exploration is extremely promising. This result, combined with the empirical observation of different alliance formation rates in different industrial sectors, or co-publication rates in scientific fields, constitutes the first step towards the empirical validation of the performance of collaboration networks.

However, we have to face the problem of a lack of precise measures to quantify the underlying knowledge regime (i.e. the agents' interaction radius) in real R&D networks, or – even worse – the lack of a consistent and reliable way to measure individual scientist trajectories in co-authorship networks. For these reasons, we will proceed in the next Chapter with an empirical analysis of a network whose agents are unequivocally locatable in a metric space, namely the R&D network, and we will take into account only the aspects of this model that can be directly tested against the data. Nevertheless, the main contribution of this model is the identification of a mechanism of volatile alliances to help the collaborating agents better explore the knowledge space.

Indeed, provided that the appropriate methodologies are known, this model paves the way for further empirical studies on collaboration networks. The scope would be to measure knowledge positions and trajectories of agents in real knowledge spaces, using – for instance – patent data for firms or publication data for scientific authors. In the case of empirical R&D networks, alliance formation rates and knowledge regimes characterizing a set of industrial sectors could be quantified and compared, allowing for a check of the consistency of our model with the observed variations in alliance activities across sectors.

Chapter 6

Toward a more general modeling framework

Summary

In this Chapter we develop an agent-based model to reproduce both the link formation and the knowledge exchange processes in a collaboration network. Based on the findings of our models on network formation and knowledge exchange, we now combine the two approaches and develop an agent based model in which agents form links based on their network features, i.e. their belonging to one of the network's circles of influence and their previous alliance history, and then exchange knowledge with their partners, thus approaching in a metric knowledge space. Furthermore, we validate the model against real data using a two-step approach. Through the SDC R&D alliance dataset, we estimate the model parameters that are related to the network, thus reproducing the topology of the resulting collaboration network. Subsequently, using the NBER data on firm patents, we estimate the parameters that are related to the knowledge exchange process, thus evaluating the rate at which firms exchange knowledge and the duration of the R&D alliances themselves. The underlying knowledge space we consider in our real example is defined by IPC patent classes, allowing for a precise quantification of every firm's knowledge position. We find that real R&D alliances have a duration of around two years, and that the subsequent knowledge exchange occurs at a very low rate. Most of the alliances, indeed, have no consequence on the partners' knowledge position: this suggests that a firm's position – evaluated through its patents – is rather a determinant than a consequence of its R&D alliances. Finally, we find that the real R&D network does not maximize the distance traveled by its agents in the underlying knowledge space. Effective policies to obtain an optimized collaboration network – as suggested by our model – would incentivize shorter R&D alliances and higher knowledge exchange rates, for instance including rewards for quick co-patenting by allied firms.

Based on M. V. Tomasello, C. J. Tessone, F. Schweitzer, "The effect of R&D collaborations on firms' technological positions", manuscript pending submission. M.V.T. took the major part in designing the microscopic rules of the model, entirely performed the computer simulations and contributed to the result analysis. The production of all plots and the writing of the manuscript were done by M.V.T.

6.1 Combining network growth and motion in a knowledge space

The agent based model that we have developed in Chapter 5 represents our first attempt to investigate a knowledge exchange process occurring in a dynamic collaboration network. That model has identified a mechanism of volatile alliances to help the collaborating agents better explore a knowledge space, using the approximation of monogamous (i.e. sparse) collaboration networks. We now introduce a more sophisticated version of such a model, to extend the validity of our previous approach to empirically observed collaboration networks, taking into account their complex and dynamic structure.

The agent-based model that we develop here constitutes the final step in the present dissertation toward a general modeling framework for collaboration networks and the knowledge exchange process occurring on top of them. Our novel agent-based model combines the realistic network formation process that we have developed in Chapter 4 with the knowledge exchange mechanisms that we have investigated in Chapter 5. The microscopic interaction rules, as well as the model validation, involve a two-step procedure that can be described as follows. The agents form links based on their network features and their social capital; the model parameters related to these mechanisms are estimated through the SDC Thomson Platinum alliance dataset. The formation of every link is then associated with a knowledge exchange process between the partners, which consequently approach in an underlying knowledge space; the model parameters related to this mechanism are estimated through firm patenting activities.

The validation of the present model is limited to the domain of R&D networks, in that they provide the most extensive and reliable data sources to test all the hypotheses on both the network topology (through alliance data) and the knowledge positions of its nodes (through patent data).

6.1.1 Social component (exploration): link formation

As mentioned in Section 6.1, the model that we develop in the present chapter combines the microscopic interaction rules of strategic link formation with those of knowledge exchange in a collaboration network. Specifically, the rules for link formation are formally identical to the ones we have presented in Chapter 4. We want every chapter of the present dissertation to be self-contained and readable independently of the other chapters. For this reason, we repeat here all these microscopic rules; for more details, refer to Chapter 4.

Node activation. We consider a network composed of N nodes; each of them is endowed with two fundamental attributes, an *activity* and a *label*. We assign to each of the $i = 1, \ldots, N$ nodes an activity a_i, that will be mapped to the empirical activities extracted from the SDC alliance dataset (Thomson-Reuters, 2013). For more details on the dataset, see Chapter 2. For more details on the calculation of the empirical activities, see Chapter 3. The activity defines the propensity of each node to be involved in a collaboration event. In particular, at every time step, a node i initiates an alliance with probability $p_i = \eta a_i \mathrm{d}t$, and the number of active nodes N_A is:

$$N_A = \eta \langle a \rangle N \mathrm{d}t, \tag{6.1}$$

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

Selection of the alliance size. When a node gets activated, it selects the number of partners m with whom the alliance is formed. We assume that the value of m is totally independent of any characteristic of the active node: we sample it, without replacement, from the empirical distribution of number of partners per alliance. In other words, we shuffle the sequence of number of partners per alliance (directly measured from the dataset) and then extract a value every time an activation event occurs; m can be thought of as the number of partners involved in every alliance event, diminished by 1, because the active node is not counted twice.

Label propagation. We assume that each of the N nodes is endowed with an attribute named *label*. This attribute is unique – i.e. every node can have only one label at any time – and fixed – once a node assumes a label, this does not change. We remember that the labels model the belonging of the agents to different groups that they implicitly define with their shared practices and/or behaviors. In the example of firms forming R&D alliances, a label symbolizes the membership of the firm in a well defined and recognized "club" or "circle of influence". In addition, we assume that such membership can be transferred to other agents as a consequence of a collaboration, provided that they are not part of any

[1] It should be noted that N_A and η slightly differ from the values we obtained in Chapter 4. This is due to the fact that the present model – as we explain in more detail in Section 6.2 – is validated on a subset of the SDC alliance dataset, precisely considering only the firms for which both the alliance and the patent data are available at the same time. This creates a bias toward larger (and thus more active) firms, causing N_A and η to slightly increase.

circle of influence yet. In our network representation, every alliance initiator does indeed propagate its label to all of its m partners, if they are non-labeled. At the beginning of every simulation, all nodes are *non-labeled*, meaning that their membership attribute is blank. There are two ways a non-labeled node can assume its label: (i) the node either receives the label from another node, if the latter initiates an alliance, or (ii) it takes an arbitrary and unique label when it becomes active for the first time (see Fig. 4.1).

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

Link formation. After deciding the category of each of its m partners, we assume that the initiator selects its specific partners within those categories according to their degree (i.e. number of previous collaborations with distinct partners). We use a linear preferential attachment rule, where the probability to attach to a node j linearly scales with its degree k_j, meaning that $\Pi(k_j) \sim k_j$. The preferential attachment rule is applied within the pool of all candidate partners, once the selection of the partner category has been made by the alliance initiator (see Fig. 4.2). This rule obviously does not apply when the initiator – be it labeled or not – decides to connect to a non-labeled node, which has by definition no previous partners ($k_j = 0$). In this case, the partner is selected among all non-labeled nodes with equal probability. When the selection process is complete, the initiator connects to its m partners. In agreement with our representation of the R&D network, we assume that all the m partners will also link to each other, forming a fully connected clique of size $m + 1$.

6.1.2 Technological component (exploitation): knowledge exchange

The second group of microscopic rules models a process of knowledge exchange between pairs of collaborating agents, similarly to what we have investigated in Chapter 5. In the present agent-based model we relax the assumption that the network can be simplified into a monogamous one (i.e. a network where every agent has only one neighbor at any point in time). We now have a network with the typical small-world and modular structure, which originates from our link formation rules and – as we have studied in Chapter 2 – is much closer to reality. Basically, we assume that every agent in the network is located in a metric knowledge space and – as a consequence of its collaborations – approaches its partners in this space. In case of multiple partners, the motion of the focal node is determined by the vectorial sum of all the effects due to each of its partners.

Location in a metric knowledge space. Every agent i is a point with coordinates $\mathbf{x_i}$, identified by a vector of D real numbers ranging from 0 to 1. In the case of R&D networks, the coordinates of every node can be thought of as the ratios of the corresponding firm's expertise along each of the D dimensions of the knowledge space. In order to validate this model against the data, we assign all agents' initial positions by using real patent data, as we explain in more detail in Section 6.2.

$$\mathbf{x}_i \equiv (x_{i1}, x_{i2}, \dots, x_{iD}) \qquad i = 1, \dots, n \tag{6.2}$$

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

$$\dot{\mathbf{x}}_i(t) = \mu \sum_{j \in \mathcal{N}_i(t)} [\mathbf{x}_j(t) - \mathbf{x}_i(t)] \tag{6.3}$$

where $\mathcal{N}_i(t)$ is the set of partners of the agent i at time t. As we can observe from Equation

6.3, in the present model there is no proximity condition for the agents' approach in the knowledge space, differently from Chapter 5. Here, the formation of the network is independent of the knowledge positions of the agents, and every link (i.e. every collaboration) has the effect to make the involved partners approach in the knowledge space. We then implement the model through computer simulations, using discrete time steps of length dt. The evolution of every agent's position $\mathbf{x_i}$ can be expressed as:

$$\mathbf{x}_i(t+\mathrm{d}t) = \mathbf{x}_i(t) + \mu \sum_{j\in\mathcal{N}_i(t)} [\mathbf{x}_j(t) - \mathbf{x}_i(t)]\,\mathrm{d}t \tag{6.4}$$

Alliance termination. Differently from Chapter 5, in the current model we do not have mechanisms such as link rewiring or interaction threshold radius, because the formation of links is determined uniquely by the network topology and the agents' attributes. However, in order to develop a more realistic model, we incorporate the termination of links as a key ingredient. We achieve this by introducing a parameter, precisely a link characteristic life time τ. We assume that the collaboration durations are distributed according to a Poisson process with rate $1/\tau$; the mean duration is obviously equal to τ. In our computer simulations, which use discrete time steps of length dt, this translates into the use of a fixed termination probability p_T for any link at any time step, equal to $p_\mathrm{T} = \mathrm{d}t/\tau$. In order to keep a simplistic set of rules, in line with our approach (see Sec. 1.2.3), we assume that the parameter τ is independent of any other feature of the network or the knowledge exchange dynamics.[2]

To sum up, in this section we have described the microscopic rules of an agent based model that is able to reproduce a dynamic link formation in a collaboration network, together with the approach of the agents in an underlying knowledge space. The learning rate μ of the agents corresponds exactly to the one we have introduced in our previous knowledge exchange model (see Chapter 5); while the link rewiring rate λ and the agents' interaction radius ε are replaced, respectively, by the link characteristic life time τ and a dynamic link formation process, described by the parameters p_s^L, p_d^L and p_{nl}^{NL} (similarly to our network formation model in Chapter 4). We summarize the model microscopic rules by means of a visual example in Fig. 6.1 and report the nomenclature of all parameters in Table 6.1.

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

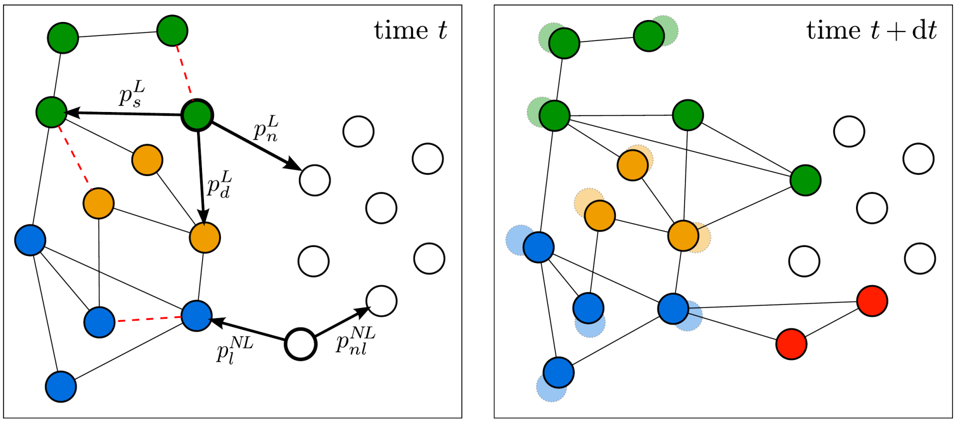


Figure 6.1: A representative example of network evolution in a bi-dimensional ($D = 2$) knowledge space. The position of the nodes in the plot corresponds to their coordinates in the knowledge space. At time $t + \mathrm{d}t$, all existing links cause the respective agents to approach in the knowledge space. Furthermore, we illustrate two collaboration events occurring at time t. The first one is initiated by a labeled node (in green), that has linked to $m = 3$ new partners, forming a fully connected clique. The second one is initiated by a non-labeled node, that has linked to $m = 2$ new partners and has taken a new arbitrary label (red). At time $t + \mathrm{d}t$, the alliance initiators propagate their labels (respectively, the green one and the red one) to the partners that were not labeled at time t yet. Finally, we illustrate the termination of 3 links (depicted with red dashed lines) at time t.

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

Table 6.1: Model parameters and their description. The "network formation" parameters are associated with the creation of new links in the collaboration network and are analogous to the ones introduced in Chapter 4. The "knowledge exchange" parameters are associated with the approach of the agents in a metric knowledge space, occurring as a consequence of a collaboration, and are similar to the ones introduced in Chapter 5.

6.2 Validation on the pooled R&D network with a two-step procedure

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

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

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

6.2.1 Alliance dataset and empirical findings

Methodology. The dataset we use to build the structure of the collaboration network is the Thomson Reuters SDC Platinum, that we already described in Chapters 2 and 4. The methodology we employ to assign nodes and links to firms and alliances, respectively, is equivalent to the one we utilized for our empirical analysis of R&D networks (see Chapter 2) and our network formation model (see Chapter 4). Again, in order to have a self-contained and independently readable chapter, we report here the main procedures and findings.

The *SDC Platinum* database (Thomson-Reuters, 2013) reports approximately 672,000 publicly announced alliances in all countries, from 1984 to 2009, with a granularity of 1 day, between several kinds of economic actors (including manufacturing firms, investors, banks and universities). We select all the alliances characterized by the "R&D" flag; after applying this filter, a total of 14,829 alliances, connecting 14,561 firms, are listed in the dataset. Furthermore, we keep in our network representation only the firms that have a corresponding entry in the patent dataset – namely the NBER dataset – that we utilize to determine their knowledge positions. This results in a network comprising 5,168 firms and 7,417 R&D alliances.

Most of the collaborations (92%) are stipulated between two partners, but some alliances – the so-called *consortia* – involve three or more partners. The distribution of the number of firms per alliance event is computed from the data and then assigned to the agents in our computer simulations, to obtain the number of selected partners m (see Section 6.1.1). The empirical distribution of number of partners that we use in this specific case is reported in Appendix E. In our network representation, we draw an undirected link connecting two nodes every time an alliance between the two corresponding firms is announced in the dataset. When an alliance involves more than two firms, we assume that all the corresponding nodes are connected in pairs, forming a fully connected clique. This choice derives from the fact that consortia, although representing only a minority of the alliances, require great coordination and resource availability from the partners. More precisely, following this procedure we obtain a total of 10,262 links from the 7,417 alliance events listed in the dataset. However, in the definition of our model, we have not made any difference between a consortium and a "standard" two-partner alliance, which is only a special case of it (and can be thought of as a fully connected clique of size 2).

We then measure the firms' *activity* distribution.[3] The activity expresses the probability that a firm takes part in any alliance event occurring in a given time window. For the validation of the present model, we use the overall firm activity, measured on the entire observation period of the dataset. We define such activity a_i, for a firm i, as the number of alliance events e_i involving firm i divided by the total number of alliance events E involving *any* firm reported in the dataset. We then assign such empirical activities a_i to the agents in our computer simulations (for the empirical activity distribution, see Appendix E).

The networks that we generate by means of computer simulations are matched to the observed R&D network with respect to three global indicators: average degree $\langle k \rangle$, average path length $\langle l \rangle$, and global clustering coefficient C,[4] which we denote as $\langle k \rangle^{OBS}$, $\langle l \rangle^{OBS}$ and C^{OBS}, respectively. The values we measure for this empirical R&D network are $\langle k \rangle^{OBS} = 3.45$, $\langle l \rangle^{OBS} = 5.05$ and $C^{OBS} = 0.11$, meaning that the network is slightly denser, more clustered, with a shorter average path length than the R&D network we analyzed in Chapter 4. As we have already found for the node activities, this happens because we now consider only the firms for which patent data are available; these firms typically have more alliance partners, thus making the network more connected.

Network formation. The approach we use to estimate the network parameters p_s^L, p_d^L and p_{nl}^{NL} is analogous to the one described in Section 4.2.1. We fix the model parameters that we can directly measure from the data, namely the number of agents $N = 5,168$, the distribution of the node activities a_i, and the distribution of number of partners m per alliance event. We stop every computer simulation when the total number of formed alliances equals the number of alliance events reported in the SDC dataset, $E = 7,417$. We vary the values of p_s^L, p_d^L and p_{nl}^{NL} in discrete steps spaced by 0.05, in the interval $(0, 1)$. The parameters p_s^L and p_d^L are bounded by the condition $p_n^L = 1 - p_s^L - p_d^L \geq 0$, meaning that their sum has to be smaller or equal to 1. This condition translates into $3,249$ points to explore in the 3-dimensional parameter space, for each of which we run 100 simulations (for a total of $324,900$ runs). We then consider the final aggregated network resulting from each of the 324,900 computer simulations and we test it against the real data with respect to three properties: average degree $\langle k \rangle$, average path length $\langle l \rangle$ and global clustering coefficient C.

We find that also in the case of the R&D network with patent data, the simulated values of $\langle k \rangle$, *meanl* and C – obtained by exploring the parameter space of the model – are

[3] For a more detailed definition and more empirical examples on agents' activity in collaboration networks see Chapter 3.

[4] For a rigorous definition of these measures, see Chapter 4.

distributed around the empirical values. However, such distributions exhibit a fairly large variance (as reported in the E). This testifies once again that our model well captures the topology of the real network for a large set of free parameters, thus allowing a meaningful exploration of the parameter space and a fine tuning of their values.

In order to identify which parameter combination is able to give the best match with the real R&D network, we use a Maximum Likelihood approach. As we have argued in Section 4.2.1, instead of having a set of observations against which we can validate our model, we only have one empirical point: the real R&D network. In particular, we cannot consider the three measures $\langle k \rangle$, $\langle l \rangle$ and C as independent, therefore the Likelihood function $\mathcal{L}$ reads as:

$$\mathcal{L}(p|net^{OBS}) = f(net^{OBS}|p) \tag{6.5}$$

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

For the choice of the error threshold ε^0, as described in Section 4.2.1, we take a conservative approach and use $\varepsilon^0 = 0.02$, that ensures a good matching with the real R&D network, without cutting out too many points in the parameters space. The corresponding Likelihood scores are reported in Fig. 6.2 by means of a 3-dimensional color map, where the color scale is representative of the Likelihood. To have a more detailed representation of the likelihood scores, we also show one slice of the parameter space obtained by fixing the parameter p_s^L to 0.45, corresponding to the highest likelihood score region, always using the error threshold $\varepsilon^0 = 0.02$. The 2-dimensional color map reported in Fig. 6.2 depict the likelihood score as a function of the other two free parameters p_d^L and p_{nl}^{NL}.

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

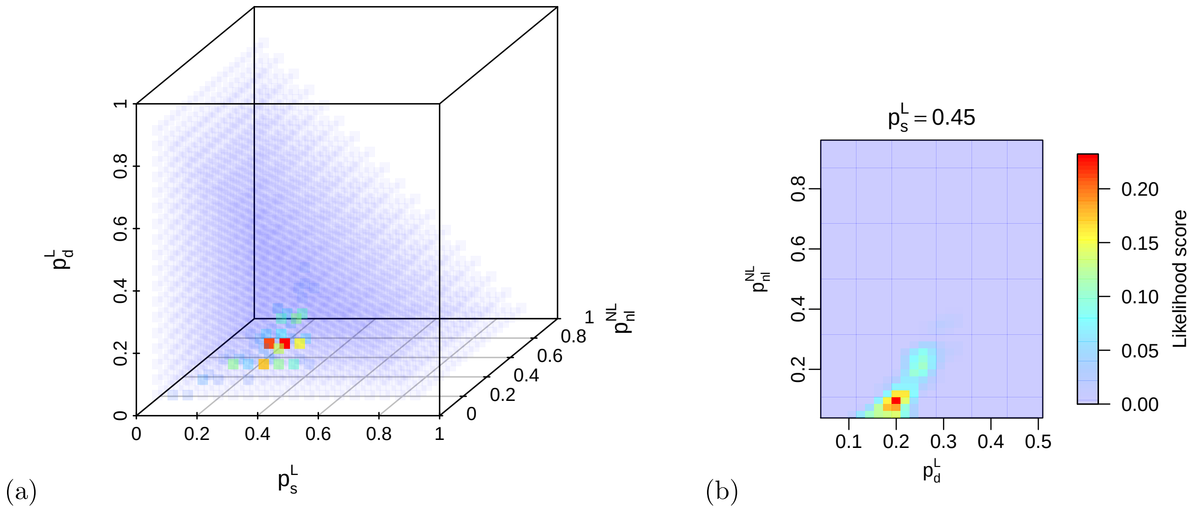


Figure 6.2: Likelihood scores for all points in the parameter space, for $\varepsilon^0 = 2\%$, represented with a 3-dimensional color map (a). After fixing the value of p_s^L to 0.45 (b), we report the Likelihood score as a function of p_d^L and p_{nl}^{NL}, using the same color scale.

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

Table 6.2: Model parameter set p^* defining the optimal simulated R&D network. The average degree, average path length and global clustering coefficient of the 100 realizations of the optimal R&D network are compared to their analogous empirical values.

probability with other non-labeled nodes ($p_{nl}^{*NL} = 0.1$). We report in Table 6.2 the set of parameter values maximizing the likelihood score, together with the values of average degree, average path length and global clustering coefficient for the simulated and the real R&D networks.

These results are in line with those we have presented in Chapter 4. However, the R&D network with patent data exhibits an even stronger tendency to favor connections with labeled nodes (i.e. incumbent firms). Due to the fact that our analysis in now restricted only to firms for which patent data are available, one could expect either an increase in the importance of network endogenous mechanisms – given that we are considering, on the one hand, larger and more active firms – or an increase in the importance of exogenous mechanisms – given that we are considering, on the other hand, firms for which the technological dimension could be more relevant in the alliance formation strategy. Our data confirm the first hypothesis, that is the increase in the relevance of network endogenous mechanisms, which results in higher probabilities for the agents to collaborate with agents that are already part of the network, and therefore already labeled. This behavior is present irrespectively of whether the alliance event is initiated by a labeled or a non-labeled node: precisely, 65% of the collaborations initiated by labeled nodes ($p_s^{*L} + p_d^{*L}$), as well as 90% of the collaborations initiated by non-labeled nodes (p_l^{*NL}), involve a labeled node as a partner.

6.2.2 Patent dataset and empirical findings

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

The approach we use to determine the knowledge position of a firm is to compute the shares of its patents in a set of different IPC classes. The first consideration has to be made on the number of classes we take into account, which will correspond to the dimensionality of the knowledge space in which the firms are located. The IPC, introduced in 1971 by the *Strasbourg Agreement*, is a hierarchical system of symbols for the classification of patents according to the different areas of technology to which they pertain.[5] A generic IPC category consists of a letter, the so-called "section symbol", followed by two digits, the so-called "class symbol", and a final letter, the "subclass". This four-character term is then followed by a group/subgroup indication, represented by additional digits. A typical IPC term can be written as follows: B34H 6/99. The sections identified by the IPC are historically stable and amount to 8, from A (human necessities) to H (electricity). The lower levels are instead subject to more frequent revisions; the eighth and last IPC edition consists of more than 120 classes, 600 subclasses, 7,000 main groups and 60,000 subgroups.

We intend to test our model on a broad set of firms, belonging to several industrial sectors, and therefore exhibiting patent activities distributed across all sections, classes and subclasses. Hence, our choice to consider only the section symbol (i.e. the first letter) in our empirical patent classification. Choosing a class- or subclass-level division would result in an excessive patent granularity, meaning a high dimensionality for the

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

corresponding knowledge space. However, for the sake of completeness, we have also tested a division at a class level (i.e. the first letter plus two digits), obtaining a total of 74 classes; we find that the computational burden of operating in a 74-dimensional space does not lead to any significant change in our results, as we show in Appendix E.

The titles of the 8 sections, as well as a patent count for each section in our dataset, is reported in Table 6.3. We find that the number of patents in all sections reflects their technological dynamism (Rosenkopf and Schilling, 2007); indeed, all sections are fairly equally represented, and the two sections exhibiting the lowest patent counts are textiles, paper and fixed constructions, two typical mature industries.

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

Table 6.3: International Patent Classification (IPC) sections and their description. The last column reports the number of patents registered in our dataset for the corresponding IPC section.

To ensure a match with our model representation, we define the knowledge position of a firm $\mathbf{x}_i \equiv (x_{iA}, x_{iB}, \dots, x_{iH})$ as the set of normalized patent counts x_{is} in each section, which in its turn equals:

$$x_{is} \equiv \frac{N_{is}}{\sum_s N_{is}} \qquad s = A, \dots, H \tag{6.6}$$

where N_{is} is the number of patents that the firm i has in a given IPC section s. In order to compute knowledge distances between pairs of firms, we use the Euclidean metric, similarly to Chapter 5. This means that the knowledge distance between two firms i and j reads as:

$$|\mathbf{x}_i - \mathbf{x}_j| = \sqrt{\sum_{s=A}^{H} (x_{is} - x_{js})^2} \tag{6.7}$$

Main empirical findings. Using the definitions provided in Eqs. 6.6 and 6.7, we now compute two empirical measures that will be later used for the validation of our model, namely (i) the knowledge positions of the 5,168 firms listed in our dataset at the beginning of the observation period – i.e. in 1984 – and (ii) the distribution of the knowledge distances between every pair of allied firms, at the moment of alliance formation. When computing the empirical knowledge position of a firm $\mathbf{x}_i$ at a given date t, we consider all the patents for which the firm has applied in a given time window Δt preceding such date t. In order to have a reliable and updated measurement, without losing at the same time too much patent information due to a short time window, we use a length equal to 5 years. We have tested different time windows, ranging from 1 to 10 years, and have found that this causes only more missing observations or noise in the distributions, with no effect on our results. The knowledge positions of the firms at the beginning of the observation period is used as an input for our computer simulations, as we explain below. In Fig. 6.3 we report the distribution of the knowledge distances between partner firms at the moment of alliance formation – from now on, the "pre-alliance knowledge distances". The minimum observed value of knowledge distance is 0, while the maximum value of knowledge distance equals $\sqrt{2}$, for normalization reasons. We find that the distribution is peaked around an intermediate distance and left-skewed, i.e. shifted toward small values. This confirms the findings that we have presented in Chapter 2, that is a preference for alliance partners to exhibit small knowledge distances. In addition, we observe that the counts drop when such preferred distance approaches zero, meaning that firms with the exact same patenting activity tend not to form alliances.

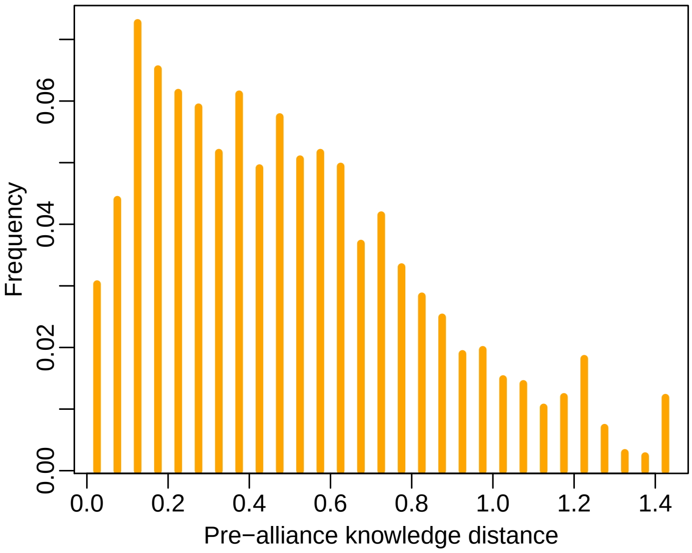


Figure 6.3: Empirical knowledge distance between every pair of partnered firms, as of the day preceding the alliance formation.

We use the aforementioned distribution to validate our agent based model and estimate the value of the knowledge exchange parameters, together with another empirical measure carrying a second, important piece of information: the distribution of the knowledge distances between every pair of allied firms, at the moment of alliance *termination* – from now on, the "post-alliance knowledge distances". However, as we have already explained in Chapter 2, the SDC dataset does not report the ending date of any alliance. To overcome this problem, during the validation of the model, we compute the empirical knowledge distance between every pair of linked firms, after a time period equal to the value of the parameter τ (in days) used in the corresponding simulation. The NBER patent dataset has a time-granularity of 1 year, thus forcing us to use a minimum 1-year time window, even when considering τ values smaller than 365 days. Nevertheless, we find that the length of such time window does not affect our results. Precisely, we find that the shape of the knowledge distance distribution appears to have the same shape, irrespectively of the time period following the alliance formation when these distances are computed, even when such time window is reduced to zero. This means that the distribution of post-alliance knowledge distances resembles the one of pre-alliance distances. In Fig. 6.4 we report the post-alliance knowledge distance distribution for different time windows of length 1, 3, 5 and 10 years.

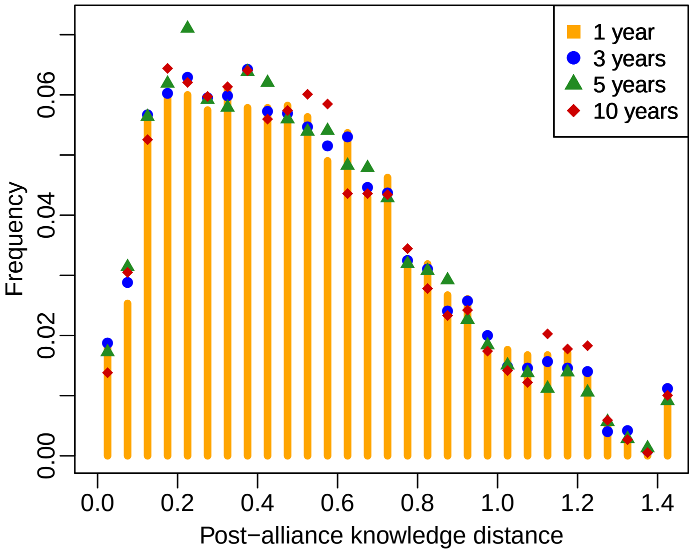


Figure 6.4: Empirical knowledge distance between every pair of partnered firms, computed 1, 3, 5 and 10 years after the date of the alliance formation.

The fact that the distribution of post-alliance distances differs only slightly from the distribution of pre-alliance distances is in agreement with another, last empirical measure we compute prior to the validation of our model. We calculate the variation of the knowledge distance separating every pair of allied firms between the moments of alliance formation and alliance termination – from now on, the “knowledge distance shift”. Again, we report our results in Fig. 6.5 for four time windows of length equal to 1, 3, 5 and 10 years. We find that the distribution of distance shifts is virtually independent of the chosen time window, as we could already expect from the distribution of the post-alliance knowledge distances. More importantly, the distribution of distance shifts is narrow and centered around zero, confirming our previous finding that post-alliance knowledge distances are subject to an overall weak change as a consequence of R&D alliances.

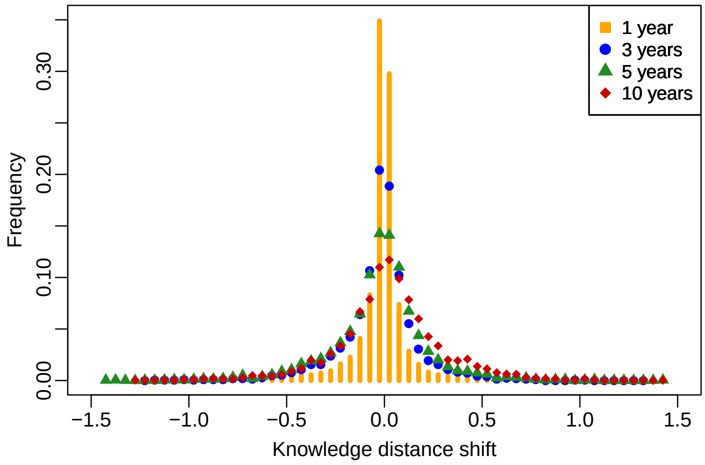


Figure 6.5: Empirical shift of knowledge distance between every pair of partnered firms, computed 1, 3, 5 and 10 years after the date of the alliance formation.

When looking at the knowledge distance shifts, we do indeed find that most of the R&D alliances cause a null change in the knowledge distance between the two partners. However, the distribution clearly exhibits tails on both sides, meaning that some alliances cause the partners to significantly move *closer* in the knowledge space, whilst some other alliances cause the partners to significantly move *farther away*. This is the result of the complex interactions between the collaborating agents and, as we show through the validation of our agent based model, it can be generated even by microscopic rules considering only an *approach* of the agents in the knowledge space, provided that this is coupled with a complex network dynamics.

6.2.3 Final model test

Exploring the knowledge exchange parameter space. We determine the values of the knowledge exchange parameters by comparing the pre-alliance and the post-alliance knowledge distance distributions in the empirical R&D network and the simulated networks generated by our model. Precisely, we fix all the network formation parameters to the values resulting from the first validation step, described in Section 6.2.1. We then fix the value of one knowledge exchange parameter that we can directly measure from the data, namely the dimensionality D of the knowledge space. As we use the eight main sections of the IPC scheme, and considering that we measure the fractions – not the numbers – of patents in each section, thus giving rise to one bounding condition, we assume $D = 7$. Consequently, the 7 numbers identifying the knowledge position of every agent are free to vary independently of each other in our simulations; the eighth component of the knowledge position can be inferred from the bounding condition that the patent fractions in every section have to sum up to 1. Obviously, each of the seven x_{is} knowledge components we use in our simulations is bound to be smaller than 1. The initial knowledge positions of the agents are assigned from the empirical data (see Section 6.2.2).

We then vary the values of the remaining knowledge exchange parameters, the agents' approaching rate μ and the characteristic alliance life time τ. We consider the values 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1 and 0.2 for the parameter μ and the values 5, 10, 20, 50, 100, 200, 300, 500, 700, 1000, 2000, 3000 and 5000 for the parameter τ, thus having a total of 104 points to explore in the parameter space. The interpretation of the parameter τ is straightforward: as explained in Section 6.1.1, we adjust the activation rate of the agents in such a way that the length of a time step $\mathrm{d}t$ can be directly interpreted as 1 day. Therefore, the value of τ, which is by design expressed in time steps, can be thought of as the characteristic duration of a real alliance in days.

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

While the pre-alliance knowledge distances are unambiguously computable on both the empirical and the simulated networks, the post-alliance knowledge distances are unambiguously computable only on the simulated networks – where every link comes to an end after a definite time. However, alliance ending dates are not available on the real R&D network. To overcome this problem, we compute the empirical knowledge distance between every pair of linked firms after a time period equal to the value of the parameter τ – in days – used in the corresponding simulation (see Section 6.1.2 for details).

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

For every simulation, we perform a two-sided KS test on the resulting pre-alliance knowledge distance distribution and the corresponding empirical distribution. We repeat the procedure for the post-alliance knowledge distance distribution, and sum the values of the two resulting $D-$statistics, thus obtaining a goodness score for every simulation. The lower such a score is, the closer the examined simulated R&D network is to the empirical one. We finally average the 100 score values for all the simulations in all points of the parameter space. Such goodness scores are presented in Fig. 6.6, where we make use of a heatmap to summarize our results.

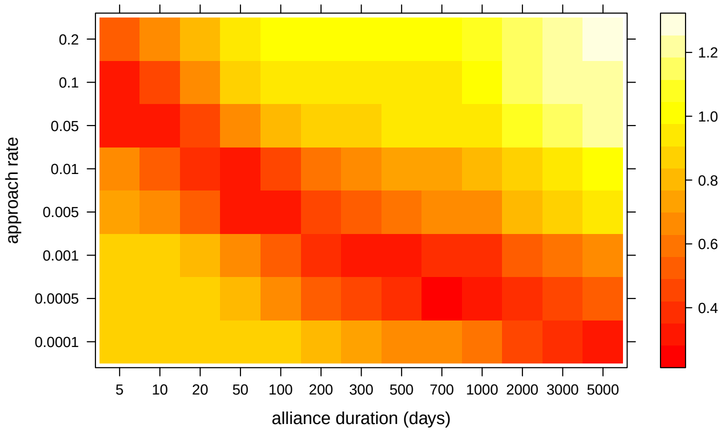


Figure 6.6: Goodness score for every point in the parameter space, depicted by means of a heatmap. The color scale corresponds to the score value; the lower the score, the closer the simulated R&D network is to the empirical one.

We find that there exists an entire region of the explored bi-dimensional parameter space maximizing the aforementioned goodness score. Such region is identified by low values of the score, corresponding to the red points in Fig. 6.6. All these points are located in the diagonal of the parameter space connecting the points having large μ values and low τ values, with those having low μ values and large τ values. This confirms our empirical finding that alliances exert a weak effect on the knowledge positions of firms.

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

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

Additional model tests. The optimal simulated R&D network, as we have shown above, is generated by the set of parameter values $\mu = 0.0005$ and $\tau = 700$. Similarly to the approach adopted in Chapter 4, we now want to investigate how well our model, fed with this optimal parameter set, is able to reproduce the knowledge distance distributions of the real R&D network. To this purpose, we report in Fig. 6.7 and Fig. 6.8 the distributions of pre-alliance and post-alliance knowledge distances, respectively. In every plot we show, the blue circles correspond to the mean values and the error bars correspond to the standard deviations of all the measures we study on the 200 realizations of the optimal simulated R&D network.

As we have imposed an equivalence criterion through the KS test, we expect that the empirical and the simulated distributions are fairly similar, which is what we find from our analysis. However, the post-alliance distance distribution generated by our model performs slightly better than the pre-alliance distance distribution. We argue that this is due to the fact that our model does not include any self-motion term for the agents in the knowledge space, as our focus is uniquely on the effect of collaborations on the agents' knowledge positions. Therefore, the pre-alliance distance distribution in our simulated network is peaked around a larger value than the real system, and then – as a consequence of the approach in the knowledge space – the post-alliance distance distribution is peaked around a slightly lower value, having a slightly better overlap with the empirical distribution.

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

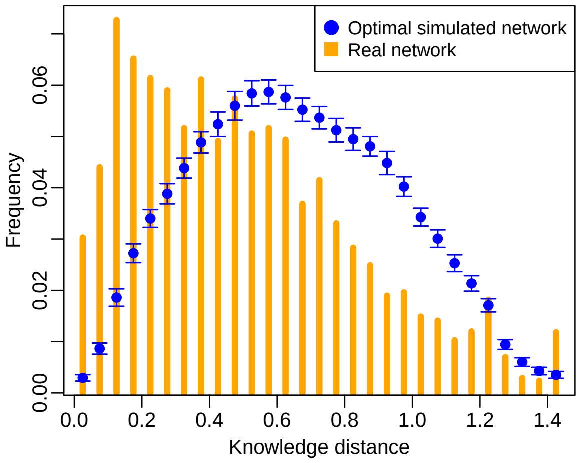


Figure 6.7: Empirical and simulated distances between firms at the moment of alliance formation.

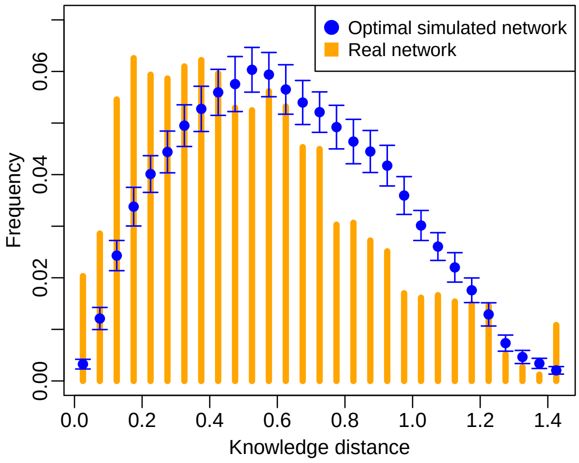


Figure 6.8: Empirical and simulated distances between firms at the moment of alliance deletion.

of collaboration formation on the positions of the agents.

Nevertheless, our model is able to reproduce one last empirical distribution – without imposing it in the validation procedure – i.e. the knowledge distance shifts. This proves that even an approach-only mechanism in a knowledge space is capable to generate positive distance shifts, i.e. increased knowledge distances between two agents as a consequence of a collaboration. We report in Fig. 6.9 both the empirical and the simulated distribution of the knowledge distance shifts for every pair of connected agents.

We find that, similarly to the real system, the simulated distance shift distribution is peaked around zero. For the reasons explained above, the collaborations in our model

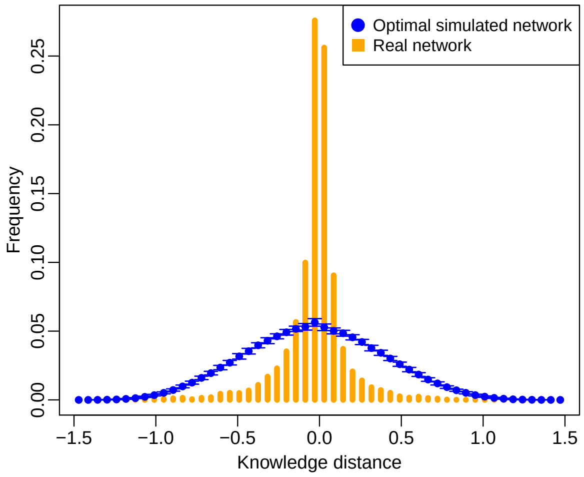


Figure 6.9: Empirical and simulated distance shifts between all allied firms.

have an overall null (or very weak) effect on the knowledge distances between agents. However, given the complex network structure characterizing the system, we also find a number of cases in which the two partners find themselves farther away in the knowledge space than they were at the moment of the collaboration establishment. Remarkably, our model can retrieve this positive right-tail of the knowledge distance shift distribution, even if the microscopic rules do not include any drift, nor self-motion, nor distancing mechanisms for the agents.

6.3 Network performance

Similarly to Chapter 5, we assume that the exploration of the knowledge space is beneficial for the whole system, and can effectively represent its own performance. In the present Section, we define such a performance indicator for our simulated networks. We do not intend to match this indicator to any possible empirical counterpart, given that we already perform our matching procedure based on empirical knowledge distance distributions. We rather want to investigate whether the empirical R&D network corresponds to a simulated network that is actually optimized with respect to this performance measure.

6.3.1 Introducing a collaboration performance indicator

Using the same notation introduced in Eq. 5.5, we define the knowledge path of an agent K_i as the sum of all the distances that the agent travels in the knowledge space during the entire simulation. Differently from the model introduced in Chapter 5, where the motion of every agent was driven by only one partner at every time step, in the present model the agents are subject to a motion resulting from interactions with multiple partners. Following this reasoning, and considering that we put the emphasis on the consequences of collaborations, we define an indicator aimed at measuring the actual effect of the collaborations in stimulating the agents' knowledge exploration. We call this indicator the *collaboration performance* $\mathcal{C}$ of the network and define it as:

$$\mathcal{C} = \int_{t=0}^{T_{\max}} \frac{N^{-1} \cdot \sum_{i=1}^{N} |\dot{\mathbf{x}}_i(t)|}{N^{-1} \cdot \sum_{i=1}^{N} k_i^{\text{act}}(t)} \, \mathrm{d}t \;=\; \int_{t=0}^{T_{\max}} \frac{\sum_{i=1}^{N} |\dot{\mathbf{x}}_i(t)|}{\sum_{i=1}^{N} k_i^{\text{act}}(t)} \, \mathrm{d}t \tag{6.8}$$

The quantity at the numerator $\sum_i |\dot{\mathbf{x}}_i(t)|$ represents the total distance traveled by all agents in the network at time t. The measure $k_i^{\text{act}}(t)$ is defined as the number of active links incident on an agent i. In this regard, we remember that not all collaborations are active at a given time t; some are terminated and become inactive, after a characteristic time τ. The quantity $k_i^{\text{act}}(t)$ measures exactly the number of active collaboration in which an agent i is involved at time t. Therefore, the ratio of the two quantities expresses the total distance traveled by the agents in the network per active link, at a given time step t, i.e. a sort of instantaneous collaboration performance of the network. This measure is then integrated over the duration $T_{\max}$ of the simulation, to obtain the overall collaboration performance $\mathcal{C}$ of the network. The quantity at the denominator of Eq. 6.8 can be thought of as the number of active links in the network at time t, which we indicate with $M^{\text{act}}(t)$,[6] multiplied by a factor 2. By plugging this into Eq. 6.8, we obtain:

$$\mathcal{C} = \int_{t=0}^{T_{\max}} \frac{\sum_{i=1}^{N} |\dot{\mathbf{x}}_i(t)|}{2 \cdot M^{\text{act}}(t)} \, \mathrm{d}t \tag{6.9}$$

We use Eq. 6.9 to compute the collaboration performance $\mathcal{C}$ in every network we generate through the exploration of our parameter space. We report our results in Fig. 6.10, by making use of a heatmap to nicely visualize the average performance $\mathcal{C}$ for every parameter combination.

6.3.2 Optimality of the real R&D network

We find that the configurations having the highest collaboration performance are located in one region of the parameter space, exhibiting high approach rates and short characteristic alliance life times. This means that an optimized network, in terms of collaboration performance $\mathcal{C}$, exhibits links with (i) a short characteristic life time and (ii) allowing for a fast knowledge transfer between the involved partners – and thus a fast approach in the knowledge space. While the dependence of the performance $\mathcal{C}$ on the approach rate μ is

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

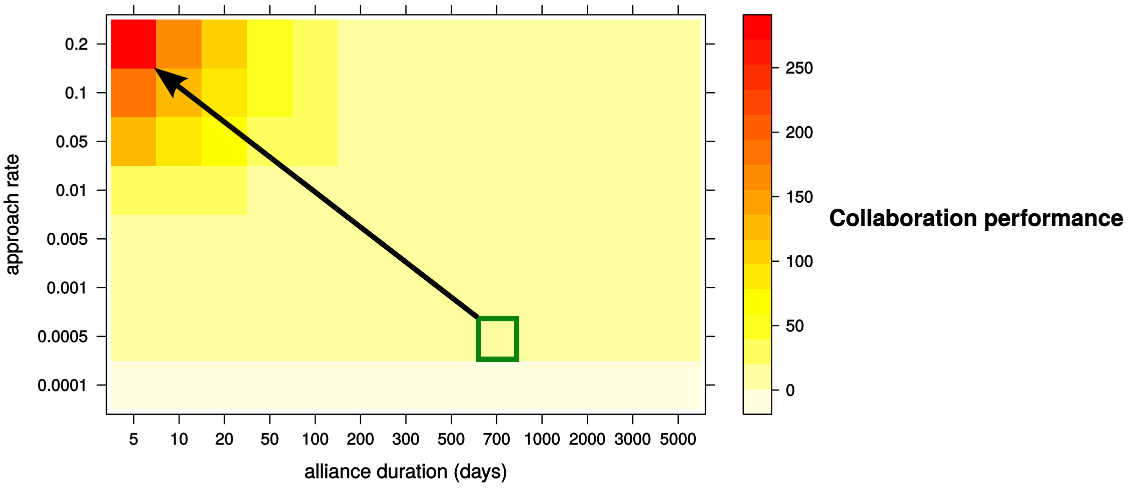


Figure 6.10: Collaboration performance of the simulated networks, as a function of the characteristic alliance duration and the approach rate. The green square in the parameter space represents the position occupied by the closest simulated networks to the real data.

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

We argue that a short collaboration life time is beneficial for the performance $\mathcal{C}$ of the collaboration network, because a reduced number of collaborations allows an agent to move efficiently along one or a few directions in the knowledge space. When the characteristic life time τ increases, more links are active at the same time, thus forcing the agents to cope with the effect of multiple partnerships; this results in a reduced motion – i.e. a reduced exploration – in the knowledge space. In other words, the density of the collaboration network increases with τ and, after a certain threshold, the addition of a new link has a negative marginal effect on the overall exploration of the knowledge space. Such non-trivial effect, which we could detect only through the implementation and development of our agent-based model, has several implications for policies aimed at optimizing real systems.

Indeed, we have found that the empirical configuration of the real R&D network is generated by parameter sets $[\mu; \tau]$ which are located along the main diagonal of the parameter space, as we show in Fig 6.6. This means that it is possible to obtain a configuration that is both *realistic* and *optimized* with respect to the collaboration performance. Therefore, effective policies to obtain an improved collaboration network would incentivize shorter R&D alliances and higher knowledge exchange rates, for instance including rewards for quick co-patenting by allied firms.

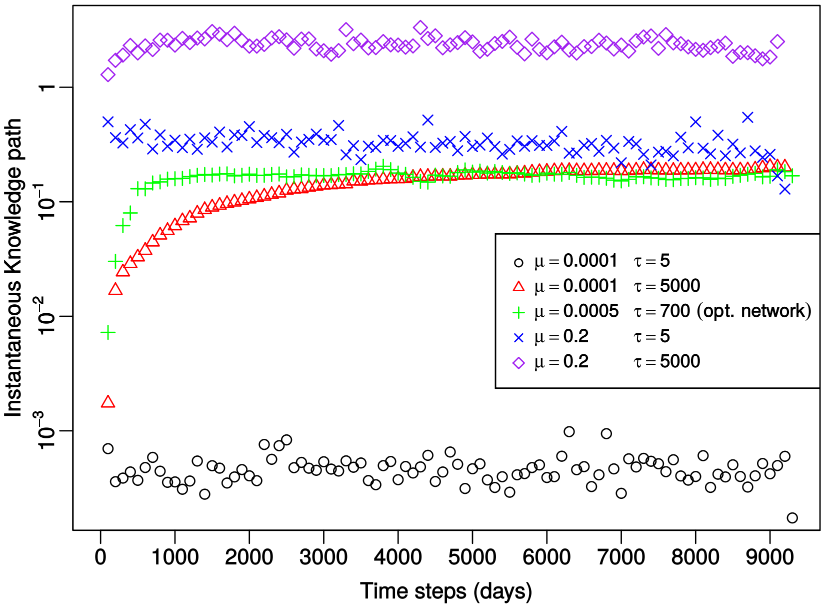


Figure 6.11: Evolution of the instantaneous knowledge path during one computer simulation. We report this quantity for 4 representative parameter sets, plus the parameter combination reproducing the real R&D network.

6.4 Discussion

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

The microscopic interaction rules are divided in two phases. In the first phase, the agents form new collaborations based on their membership attribute, i.e. their label. Similarly to our network formation model introduced in Chapter 4, Such attribute can be propagated to other agents as a consequence of an alliance, thus defining the so called *circles of influence* (groups of agents sharing the same membership attribute). The model includes different link formation probabilities depending on both the alliance initiator's and its future partners' membership attributes. In the second phase, all pairs of connected agents exchange knowledge and approach each other in a D-dimensional knowledge space, with a characteristic rate μ. The collaborations have a characteristic life time τ; after a collaboration is terminated, the approach of the involved partners ceases. The linking probabilities constitute the *network formation* parameters, while the approach rate μ, the collaboration characteristic life time τ and the dimensionality D of the knowledge space represent the *knowledge exchange* parameters of our model.

The validation of our model against real data has been performed through a novel two-step approach as well. By means of the SDC alliance dataset, we estimate the network formation parameters, thus reproducing the topology of the resulting collaboration network. Subsequently, through the NBER dataset (on firm patents), we estimate the knowledge exchange parameters, thus evaluating the rate at which firms exchange knowledge and the duration of the R&D alliances themselves. The underlying knowledge space we consider in our real example is defined by IPC patent classes, allowing for a precise quantification of every firm position.

By running extensive computer simulations, we have identified the set of linking probabilities generating the closest network to the real R&D network, with respect to average degree, global clustering coefficient and average path length. As summarized in Table 6.2, when the initiator of the alliance is a labeled node (i.e. an incumbent firm), it connects to a node having the same label with probability $p_s^L = 0.45$, to a node having a different label with probability $p_d^L = 0.2$ and, consequently, to a non-labeled node (i.e. a newcomer firm) with probability $p_n^L = 0.35$. When the alliance is initiated by a non-labeled node (a newcomer), it connects to a labeled node with probability $p_l^{*NL} = 0.9$ or to another non-labeled node with probability $p_{nl}^{*NL} = 0.1$.

These results are in line with those we have presented in Chapter 4. However, the R&D network with patent data exhibits an even stronger tendency to favor connections with labeled nodes (i.e. incumbent firms). Due to the fact that our analysis in now restricted to firms for which patent data are available, we find an increase in the importance of network endogenous mechanisms, given that we are considering larger – and therefore more network-active – firms. The tendency to connect to labeled nodes is present irrespectively of whether the alliance event is initiated by a labeled or a non-labeled node: precisely, 65% of the collaborations initiated by labeled nodes ($p_s^{*L} + p_d^{*L}$), as well as a surprising 90% of the collaborations initiated by non-labeled nodes (p_l^{*NL}), involve a labeled node as a partner. In this regard, the validation of our model brings additional support to the theory of the importance of existing network structures in the formation of new collaborations.

As for the knowledge exchange parameters, we find that the real R&D network is best reproduced by a configuration exhibiting a relatively low approach rate ($\mu = 0.0005$) and a characteristic duration of around two years ($\tau = 700$ days). Both the test of our agent based model and our empirical analysis, indeed, show that collaborations exert an overall weak effect on the partners' knowledge position. However, by examining the distribution of the knowledge distance shifts between every pair of connected agents, we find both a positive and a negative tail. Despite the overall weak effect, some collaborations can cause extreme shifts: some bring the partners closer, while some others push them farther in the metric knowledge space.

The finding of a typical life time τ of around 2 years is consistent with our previous theoretical assumptions and a number of previous studies (see Chapter 2). It should be noted that the real R&D network can be reproduced by a whole set of parameter combinations, lying on the main diagonal of the parameter space formed by μ and τ. Precisely, these points exhibit large μ values and low τ values, or low μ values and large τ values, thus confirming the weak effect of real collaborations on knowledge positions: the faster the approach rate is, the shorter the characteristic alliance life time has to be to generate a system corresponding to reality, and vice-versa.

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

Finally, we have investigated the performance of our generated collaboration networks with respect to a new performance indicator. We call such indicator the collaboration performance $\mathcal{C}$ of the network, and define it as the distance traveled by all agents per active link (we dynamically compute this measure at every time step and then perform a complete exploration of the parameter space in order to find the optimal network). We find that the configuration exhibiting the highest performance $\mathcal{C}$ has the shortest possible characteristic alliance duration τ, and the largest possible approach rate μ. This new, non-trivial result has some implications for policies aimed at optimizing real systems.

Indeed, we have found that the real R&D network is generated by parameter sets $[\mu;\, \tau]$ which are located along the main diagonal of the parameter space. This means that it is possible to obtain a configuration that is both *realistic* and *optimized* with respect to the collaboration performance. In the case of real R&D alliances, obviously, it would be impossible to require alliance durations as short as 5 or 10 days; moreover, it is not easy to directly enforce a fast approach or learning rate between real companies. However, the results of our simulations suggest that effective policies to obtain a *collaboration-performing* network would incentivize shorter R&D alliances and higher knowledge exchange rates. Such policies could include, for instance, rewards for co-patenting activities from partner companies, when these are carried out the earliest possible after the establishment of an R&D alliance. The goal is to push companies to always explore new knowledge positions with new partners, although limiting the duration of a single alliance, and avoiding having too many active collaborations at the same time.

To sum up, our model can successfully reproduce the network topology and the distribution of the agents' knowledge positions in a real collaboration network, while providing at the same time a unique methodology to estimate an indicator of network performance. Although we limit its validation to the domain of R&D networks, we argue that our model is flexible and extendable to other collaboration networks, whose nodes can be unequivocally positioned in a knowledge space. For the moment, we have been forced to skip the validation on co-authorship networks, because of the lack of a clear methodology to locate their nodes (i.e. the authors) in a knowledge space – unfortunately, the typical knowledge classifications are instead applied to the links of the network (i.e. the papers).

In conclusion, we argue that – to the best of our knowledge – the novel modeling framework we have developed in the present Chapter offers the most complete and straightforward interpretation of the effects of knowledge exchange in a dynamically evolving collaboration network.

Chapter 7

Discussion and conclusions

7.1 Our contributions

In this thesis we have studied the formation and evolution of collaboration networks using a *complex systems* perspective. Precisely, we have focused on those collaboration networks in which every link formation event involves a knowledge flow. We have started our analysis by selecting two prominent examples of such systems, namely R&D and co-authorship networks, and thoroughly studied them from an empirical point of view.

Next, we set the goal to understand the microscopic rules leading to link formation and dissolution between individual agents, how they affect the aggregate performance generated by these systems, and whether it is possible to optimize real systems with respect to such performance. All of these questions have been addressed by means of agent-based models. Our findings can be summarized along two lines: purely empirical and model-driven.

7.1.1 Empirical findings

Chapter 2. The analysis carried out in Chapter 2 on R&D networks has several implications. Rather than focusing on sectoral-related differences, our results provide strong support to the hypothesis that many R&D network properties are robust across several manufacturing and service sectors. These properties are also invariant across two different scales of aggregation. That is, they are the same if one considers the R&D alliances irrespectively of the sectors to which the firms belong (pooled network), or if one considers only alliances centered on a sector (sectoral networks). These properties span from basic network characteristics such as size, density, degree distributions, to more complex features such as the presence of small worlds and core-periphery architectures.

Remarkably, this reflects the similarities present at the microscopic level in the rules determining alliance formation. Through our econometric model in 2.4, we have shown that alliance preferences between firms with small geographical, sectoral, technological and network distances are stable and robust across sectors.

Nevertheless, our results also show that not *all* properties of the network are invariant across different scales of aggregation. We have found that sectoral R&D networks are disassortative, i.e. characterized by a negative correlation across node degrees, whereas the pooled network is assortative. This transition from disassortative to assortative networks is a new fresh stylized fact, that we have not further investigated in this dissertation, but that should be taken into account in the theoretical explanations of R&D networks.

This instability of degree-degree correlations in R&D networks is reflected in our findings at the microscopic level as well. While in the pooled R&D network the firms most likely to form an alliance exhibit a strong centrality disparity, in sectoral R&D networks this tendency disappears to be replaced by sectoral specific behaviors. This finding also stresses the importance of using agent-based models to better understand and interpret certain stylized facts. Despite the fact that firms with a high centrality disparity (i.e. showing local disassortativity) are more likely to form new links in the network, the resulting pooled R&D network is generally assortative. As already mentioned, we leave the study of assortativity and disassortativity in collaboration networks to our future research; however, this phenomenon and its emerging consequences deserve further attention.

Next, the result that both the pooled and sectoral networks are organized into core-periphery architectures – nested structures in particular – supports the predictions of the recent theoretical literature (e.g. Goyal and Joshi, 2003; Westbrock, 2010), and more precisely of the knowledge-recombination model of König *et al.* (2012). In this model, in case of relatively costly partnerships, the resulting efficient R&D network exhibits the nested structure that we observe empirically.

These findings are further supported by our econometric approach, which shows – interestingly – that alliances are more likely to be observed if they maximize the change in some aggregate network measures, i.e. eigenvalue of the connected component to which the firms belong (König *et al.*, 2012) or the harmonic average path length of the network (Jackson and Wolinsky, 1996, e.g.), rather the increase of the single firm centralities. The network topology resulting from this behavior, again, is compatible with the observed nested architectures. Most likely, this does not mean that firms are not concerned with the improvement of their own network centrality when establishing new alliances. We argue instead that firms do try to maximize their expected return, but this may depend on both network-related and network-unrelated factors. The complex interdependencies between firm decisions and the actual alliance formation give rise to a network growth process where the newly formed links tend to maximize some aggregate network indicators rather than individual firm centralities. The result is the observed coefficient sign in our econometric model.

Similarly, the data do not show any evidence of another theoretical aspect, i.e. costly R&D alliances: firm dyads engaged in more distinct alliances are more likely to form one additional alliance themselves, thus not showing any limitation effect in the number of newly established R&D alliances. Again, this might be an effect of the complex interdependencies above mentioned; only the use of an agent based model – that we investigate in the next chapters – will be able to give us further insights.

Another important empirical finding is that *previous network structures*, along with potential network structure changes, *matter in the alliance formation*, as testified by the broad and right-skewed degree distribution in all R&D networks, as well as by our econometric approach. Even though the network-unrelated variables alone have a slightly better predictive power than network-related variables alone, we have shown that a model including both types of variables has the highest possible goodness of fit when explaining the formation of R&D alliances. In addition, the analysis of the predictor coefficients allows us to identify an additional set of invariant and sectoral-robust features at the microscopic level. Namely, alliances are more likely to be established if the potential partners belong to the same country and sector, and exhibit a small technological distance; if they have already engaged in many alliances with other distinct firms; if they are already – directly or indirectly – connected by a path in the R&D network; if the formation of the considered link leads to a small increase in the individual firm centralities, but a large increase in a set of aggregate network centrality indicators, as already discussed.

Going further, our results show that the rise and fall of R&D networks has been mainly driven by the entry and exit of firms participating into alliance activities (see Section 2.3.1). Our interpretation is that the sheer number of firms participating in R&D alliances – as well as the number of scientists writing papers, to mention our next example – is an *exogenous* factor with respect to the network. Moreover, the entry/exit dynamics of agents plays a significant role in the network formation and evolution. This means that the observed rise-and-fall trends in R&D networks are a consequence of the exceptional firm activity in the mid-nineties, fueled by the IT-bubble and subsequently continued by the bio-tech revolution; this is in agreement with Schilling (2009), that has detected the same mid-nineties peak in several alliance dataset, also across countries and sectors.

However, this does not change the relevance of our results. The fact that such rise-and-fall trend characterizes many network properties means that the collaboration network organizes itself in a very peculiar way, which is the true object of our study. Therefore, our analysis answers the question: given certain exogenous factors that cause more or fewer firms to be part of a network, how do they *self-organize* their collaborations? And what is the structure of the emerging network?

The answer is that R&D networks are able to self-organize into components having complex characteristics, namely small-world, core-periphery and nested architectures. And these structures emerge as a consequence of the microscopic behaviors that we have tried to identify with our econometric model – mainly, tendency to close network paths and to maximize the overall network connectedness. In this respect, we extend the results of Gulati *et al.* (2012) – that has observed one of the complex features (small world properties) in one sector (computer industry) – to many industrial sectors, not limited to manufacturing.

Chapter 3. In Chapter 3 we have extended the investigation of network trends and patterns from R&D to co-authorship networks in scientific disciplines. We find that co-authorship networks are characterized by similar network structures to the R&D networks, i.e. the emergence of giant connected components, heterogeneous degree distributions and small world properties. Differently from R&D networks, co-authorship networks are characterized by a positive degree assortativity coefficient and do not exhibit any rise-and-fall trend. On the contrary, they are characterized by generally rising trends over the last three decades, associated with fluctuating trends for degree heterogeneity across nodes, assortativity and small world properties.

As already mentioned for the R&D networks, we believe that the unprecedented growth that has characterized every scientific field – and the corresponding publication rates – in the recent years is an *exogenous* event with respect to the network formation and evolution. Therefore, we do not aim at explaining this factor. However, the emergence of similar network features (i.e. heterogeneous and right-skewed degree distribution, assortativity and small world properties) suggests the existence of some invariant mechanisms, albeit associated with domain-related specificities, determining the self-organization and the evolution of collaboration networks.

Considering that our aim is to identify the minimal set of microscopic rules able to reproduce the topology of such networks, we have investigated a different set of features, more elementary and primitive than the ones studied in Chapter 2, thus representing more suitable basic blocks for an agent-based model. The features that we have studied are: i. the size of collaboration events (i.e. firms per alliance or authors per paper), ii. the agents' *activity* (i.e. their propensity to engage in a collaboration) and iii. structural communities in the network (beyond the agents' sectoral or geographical positions).

The distribution of agents per collaboration is broad and right-skewed for all R&D and co-authorship networks, even though the co-authorship networks exhibit a higher degree of variability across fields. The number of agents per collaboration event spans from 2 (the vast majority in all networks) to 55 (in the relativity and gravitation co-authorship network). The agents' activities distribution are dispersed and right-skewed as well, spanning several orders of magnitude. Differently from many networks indicators, the activities are stable and can effectively model the propensity of every agent to engage in a collaboration event, thus making them viable candidates for an agent attribute in our model. In addition, this study represents the first example of empirical activity computation in an economic network.

Next, we have detected the presence of modular structures in all collaboration networks, through a well-known community detection algorithm (Infomap). Such finding is significant and robust across domains. However, being it a complex and emerging topological property of the network, we decide to use it as a validity test, and not as a building block of our model. Such network clusters are a network-topology-based feature, and are not explained by the belonging of the agents to same country or sector: we rather argue that the modular structures are indicative of some microscopic rules of strategic link formation, that involve the presence of a latent membership attribute. The existence of this attribute, together with specific rules of propagation during the establishment of collaborations, is at the basis of our the agent-based model that we develop in Chapter 4.

7.1.2 Model-driven findings

Chapter 4. Inspired by our empirical findings, especially on R&D networks, we have designed a model where the agents, representing real collaborating agents, are endowed with two key attributes: an activity (representing their propensity to engage in new alliances) and a label (representing their membership in a given circle of influence).

The simple yet effective set of microscopic rules that we have proposed includes both network-endogenous and network-exogenous mechanisms for link formation. Our model is centered around the assumption that the agents have a membership attribute, that we call *label*. Such attribute can be propagated to other agents as a consequence of a collaboration, thus defining the so called *circles of influence* (groups of nodes sharing the same membership attribute). The model includes different link formation probabilities, that depend on both the collaboration initiator's and its partners' membership attributes.

We have first tested our model against the SDC Platinum alliance dataset. By running extensive computer simulations, we have identified the set of linking probabilities that generates the closest network to the empirical pooled R&D network, with respect to average degree, global clustering coefficient and average path length. We have found that a labeled node (i.e. an incumbent firm) connects to a node having the same label with probability $p_s^{*L} = 0.3$, to a node having a different label with probability $p_d^{*L} = 0.3$ and, consequently, to a non-labeled node (i.e. a newcomer firm) with probability $p_n^{*L} = 0.4$. A non-labeled node (a newcomer), when initiating a collaboration, connects instead to a labeled node with probability $p_l^{*NL} = 0.75$ and to another non-labeled node with probability $p_{nl}^{*NL} = 0.25$. The optimal simulated network generated by our model exhibits network measures that deviate from the empirical values by less than 2%.

Given that 60% of the alliances initiated by incumbents, and 75% of the alliances initiated by newcomers, are directed towards incumbents, we confirm the importance of the endogenous mechanisms over the exogenous ones in the formation of new collaborations. This result is confirmed in all sectoral R&D networks and all co-authorship networks: alliances initiated by incumbents are preferably directed towards other incumbents, with a preference for nodes in the same circle of influence. However, in the totality of co-authorship networks, newcomers tend instead to form their first alliance with other newcomer nodes.

Next, we have performed further tests to check whether the model is able to reproduce a set of microscopic network properties, even without imposing any equivalence in the validation procedure. For all examined collaboration networks, we have obtained a surprising agreement with the empirical data. Our model, fed with the optimal parameter combinations, is able to reproduce the distributions of degrees, path lengths, local clustering coefficients and network component sizes. We have also retrieved the distribution of path lengths between every pair of nodes at the moment of link formation, especially including the counts for path lengths 1 (i.e. repeated collaborations) and 2 (i.e. triadic closures). This strongly supports the goodness of our model microscopic rules.

In addition, we have found a remarkable overlap between the network partition defined by a widely used community detection algorithm (Infomap) and the one defined by our node labels (i.e. membership attributes). Such overlap, measured through a normalized mutual information criterion, is around 90% for all collaboration networks. We argue that our label propagation mechanism models the formation of network clusters in an efficient fashion: this result is remarkable if we consider that the Infomap algorithm detects structural clusters based on the probability flow of random walks in the network (Rosvall and Bergstrom, 2008), while our label propagation mechanism consists of an assignment of a fixed membership attribute – which is easier to map to a real phenomenon.

Chapter 5. We have developed an agent based model of dynamic collaboration formation and knowledge exchange, and introduced a novel network *performance measure*. By studying the interactions of the agents in a *metric knowledge space*, and assuming that collaborations bring agents closer in this space, we have found that the system follows a non-trivial dynamics and reaches a steady state in which the agents cluster around a set of emerging attractors.

The two model parameters we vary are the interaction radius between the agents ε and the alliance rewiring rate λ in the network. We have found that the number of knowledge clusters decreases by increasing the threshold interaction radius ε, because the agents are able to collaborate with partners located farther away in the knowledge space, thus converging all together towards one position. When the knowledge regime is strongly tacit or strongly explicit, the number of knowledge clusters depends only on ε itself, and not on the alliance rewiring rate λ. The most interesting case occurs for intermediate knowledge regimes, in which the number of knowledge clusters increases with λ. Small rewiring rates lead to the emergence of only one knowledge cluster, which is dispersed in the knowledge space and does not clearly exhibit the presence of a knowledge attractor in it. A faster alliances rewiring leads the agents to potentially have collaborations with more partners, allowing the emergence of a larger number of knowledge clusters; in this case, the presence of knowledge attractors, around which the firms eventually cluster, is clearly visible.

More importantly, we have found that the nodes, as a result of their interactions, explore a certain distance in the knowledge space $\langle K \rangle$, which we consider as the *performance measure* of the collaboration network. We have found that there exists a specific parameter combination maximizing such indicator of performance, specifically intermediate values of both the rewiring rate λ and the interaction radius ε. In particular, if we focus on the dependence of $\langle K \rangle$ on λ, given a fixed ε, we find that there exists an optimal value λ^* maximizing $\langle K \rangle$. Such optimal rewiring rate λ^* exhibits a weak dependence on ε; namely, it slightly decreases when ε increases (but only for intermediate values of ε). This is consistent with the empirical study by Rosenkopf and Schilling (2007), that shows a varying alliance formation rate across industrial sectors.

Similarly, we have found that, given a fixed alliance rewiring rate, there exists an optimal interaction radius ε^* maximizing the mean knowledge path $\langle K \rangle$. In conclusion, we have identified through this model a mechanism of volatile alliances to help the collaborating agents better explore the knowledge space.

Chapter 6. We have developed an agent-based model that is able to reproduce both the link formation and the knowledge exchange processes in a collaboration network, by combining our previous agent-based models. In this new modeling framework, agents form links based on their network features and then exchange knowledge with their partners, thus approaching in a metric knowledge space. Our agents are endowed with three key attributes: an *activity* (representing their propensity to engage in new alliances), a *label* (representing their membership in a given circle of influence), and a *position in a metric knowledge space* defined by a vector (which can be thought of as the fractions of the agent's knowledge in some categories).

The microscopic interaction rules are divided in two phases. In the first phase, the agents form new collaborations based on their membership attribute, i.e. their label, following the same rules as our network formation model, explained in Chapter 4, including different link formation probabilities depending on both the alliance initiator's and its future partners' membership attributes. In the second phase, all pairs of connected agents exchange knowledge and approach each other in a D-dimensional knowledge space, with a characteristic rate μ. The collaborations have a characteristic life time τ; after a collaboration is terminated, the approach of the involved partners ceases. The linking probabilities constitute the *network formation* parameters, while the approach rate μ, the collaboration characteristic life time τ and the dimensionality D of the knowledge space represent the *knowledge exchange* parameters of our model.

We estimate the network formation parameters through the SDC alliance dataset and the knowledge exchange parameters through the NBER dataset on firm patents, thus quantifying the rate at which firms exchange knowledge and the duration of the R&D

alliances themselves. The underlying knowledge space we consider in our real example is defined by IPC patent classes, allowing for a precise quantification of every firm position.

The results we have found for the network formation parameters are in line with those of Chapter 4. The R&D network with patent data exhibits an even stronger tendency to favor connections with labeled nodes (i.e. incumbent firms), irrespectively of whether the alliance event is initiated by a labeled or a non-labeled node. Precisely, 65% of the collaborations initiated by incumbents, as well as a surprising 90% of the collaborations initiated by newcomers, involve an incumbent as a partner. In this regard, the validation of our model brings additional support to the theory of the importance of existing network structures in the formation of new collaborations, even in a collaboration network where the technological positions of the agents play an important role.

As for the knowledge exchange parameters, we find that the real R&D network is best reproduced by a configuration exhibiting a relatively low approach rate, $\mu = 0.0005$. Both the test of our agent based model and our empirical analysis, indeed, show that collaborations have an overall weak effect on the partners' knowledge position: this suggests that a firm's position – evaluated through its patents – is rather a *determinant* than a consequence of its R&D alliances, in agreement with Sampson (2007). However, by examining the distribution of the knowledge distance shifts between every pair of connected agents, we find both a positive and a negative tail. This indicates that some R&D collaborations can have extreme effects on the distance between the involved agents, by bringing them much closer or much farther in the knowledge space.

We have then found that the typical life time τ for an R&D alliance is around 2 years, precisely 700 days. This is consistent with our previous theoretical assumptions and a number of previous studies (see Chapter 2), e.g. Phelps (2003). It should be mentioned that the real R&D network can actually be reproduced by a whole set of parameter combinations, lying on the main diagonal of the parameter space formed by μ and τ. Precisely, we refer to those points having large μ values and low τ values, or low μ values and large τ values. Again, this is consistent with the finding that collaborations have a weak effect on the agents' knowledge positions.

Finally, we have investigated our simulated collaboration networks with respect to a new indicator, that we call the *collaboration performance* $\mathcal{C}$. We define it as the distance traveled by all agents, divided by the number of active links in the network, dynamically computed at each time step. The configuration exhibiting the highest performance $\mathcal{C}$ has the shortest possible characteristic alliance duration τ, and the largest possible approach rate μ. This new, non-trivial result has some implications for policies aimed at optimizing real systems. Indeed, considering that the real R&D network is generated by parameter sets having high μ and low τ, or vice-versa, this means that the system can be steered towards a configuration that is both *realistic* and *optimized* with respect to the collaboration performance, with short alliance duration and high knowledge exchange rate.

7.2 Applications

From a general empirical perspective, our results generalize previous findings in the literature, that were limited to the analysis of few industrial sectors or scientific research fields. From a theoretical perspective, the fact that many properties of collaboration networks hold irrespectively of the domain, or the scale of aggregation, opens up the fascinating possibility that the same universal mechanisms are responsible for the emergence of those features.

Indeed, we believe that the major contribution of our study is to provide a straightforward and universal methodology to study collaboration networks, while at the same time assessing their performance.

7.2.1 Methodology to systematically characterize networks

The use of the agent-based models that we have developed in Chapters 4 and 6 can be in principle extended to any collaboration system for which a series of time-stamped links or alliances are available. We argue that the predictive power of our model lies in the simplicity of its label propagation rules and the flexibility of its linking probabilities. Moreover, the agent *activity* encodes in an effective and simple fashion a big deal of information on the system under examination, including the naturally heterogeneous propensity of the agents to engage in new collaborations, and the entry of newcomers in the network (a first-time agent activation can be equally considered as an entry to the network).

The linking probabilities deriving from the subsequent validation procedure on the dataset at hand can give precise insights into the strategies pursued by the collaborating agents. For instance, if we take the linking probabilities estimated for the pooled R&D network, our findings suggest that incumbent firms tend to have a preference towards other incumbent firms: 60% of their alliances belong to this category, split between a 30% probability to connect to a node in the same circle of influence and a 30% probability to connect to a node in a different circle of influence. In the remaining 40% of the cases, incumbents form alliances with newcomers: these alliances are driven only by exogenous factors, since there cannot be any network endogeneity affecting nodes that are not part of the network yet.

On the other hand, newcomers link to incumbent firms in 75% of the cases. Such alliances are driven by network endogenous factors, namely the newcomers' motivation to join the R&D network by partnering with firms that are already part of it. However, a fraction (25%) of alliances initiated by newcomers are directed to other newcomers. The reasons behind these alliances are related to exogenous factors such as the firms' commercial or technological capital. Some newcomers prefer to join the R&D network by partnering with other newcomers with no network experience – for instance, small start-up companies in highly technologically dynamic environments – rather than with an incumbent firm.

The extended validation procedure on sectoral R&D networks as well as co-authorship networks has given us further insights, which are very informative about the nature of the system under examination. Fox example, in both R&D and co-authorship networks, labeled nodes (incumbents) tend to form links with other labeled nodes. Besides, when forming a link with another labeled node, the collaboration initiator tends to select a node having the same label, i.e. belonging to the same circle of influence; this tendency is less pronounced in the pooled R&D network and the sectoral R&D networks characterized by high technological dynamism, where incumbents exhibit a balanced alliance strategy, and is instead much stronger in the totality of co-authorship networks, where the circles of influence drive the formation of links between incumbents. In all co-authorship networks, plus the sectoral network of R&D, laboratory and testing (again, a highly technologically dynamic sector), non-labeled nodes – i.e. newcomers – tend to form their first links with other non-labeled nodes. For the rest the sectoral R&D networks, instead, non-labeled nodes (newcomers) tend to enter the network by forming a link with labeled nodes, i.e. incumbents.

The last finding, we argue, highlights the existence of higher entry barriers in economic systems than in academic environments. This result is consistent with empirical evidence: unlike newcomer firms, which join the R&D network for the first time by partnering with incumbent firms, a young scientist writes his/her first paper mostly with other young scientists, and only a small part of the co-authors are expert scientists (typically, one post-doctoral researcher or the professor in the same group).

So far, the fine tuning of our model has suggested that, in most cases, endogenous mechanisms for network formation are predominant over the exogenous ones, or – in other words – the existing network structures explain most of the newly formed links. We envision a broad range of application for this model on several collaboration networks, possibly in other domains such as open source software projects, online social networks, political networks; the model validation and fine tuning could extending our current findings or disprove them in some of the examined systems.

7.2.2 Design of optimal collaboration networks

The more general modeling framework that we have developed in Chapter 6 not only could successfully reproduce the topology and the distribution of the agents' knowledge positions in a real collaboration network, but could also provide a unique methodology to estimate an indicator of network performance. However, this leaves some open questions that need to be further investigated. While the network topology can be unequivocally matched between real data and simulations, the network *performance* involves a more arbitrary definition.

In the case of the pooled R&D network with patents, we have used a measure of knowledge exploration as performance indicator. Precisely, we have defined a collaboration performance $\mathcal{C}$ defined as the distance traveled by all agents in the knowledge space, divided by the number of active links. This definition reflects the concepts presented in Cowan and Jonard (2004); however, it could be argued that other measures could give different (if not better) indications. For instance, one could employ the sheer number of patents (or publications) in the computation of the performance (Jaffe and Trajtenberg, 2002), or a uniquely network-related measure, such as some aggregate connectedness measure (König *et al.*, 2012, e.g.).

The results based on our collaboration performance $\mathcal{C}$ suggest that effective policies to improve a real R&D network would incentivize shorter R&D alliances and higher knowledge exchange rates. Such policies could include, for instance, rewards for co-patenting activities from partner companies, when these are carried out the earliest possible after the establishment of an R&D alliance. The goal is to encourage companies to always explore knowledge positions with new partners, although limiting the duration of a single alliance, and avoiding having too many active collaborations at the same time.

We believe that a knowledge exploration-related indicator captures in a better and less biased way the performance of such a system, because it takes into account the real knowledge trajectories of the agents. The sheer number of patents and citations for R&D networks – or, even worse, the number of publications and citations in co-authorship networks – would exhibit a strong time dependence and does not necessarily reflect the goodness of the produced knowledge.

However, we do envision the inclusion of network-related measures to better capture the knowledge diffusion properties of the collaboration systems. In agreement with the models proposed by König *et al.* (2012) or Jackson and Wolinsky (1996), we believe that a higher network connectedness is beneficial for a faster knowledge diffusion. In this regard, some preliminary results — that we do not show in this dissertation — indicate that the knowledge diffusion speed increases with the small world properties of the collaboration network. If this were confirmed by further analyses, it would mean that we can improve the performance of the network by tuning the microscopic parameters in such a way that its average path length decreases and its clustering coefficient increases.

In terms of linking probabilities (see Chapters 4 and 6), this would correspond to incentivize alliances between incumbents in the same circle of influence, at the expense of nodes from different circles. At the same time, the newcomer nodes should form more links with each other, with a small part of collaborations directed to incumbent nodes. Remarkably, these strategic linking probabilities resemble the ones that we have observed in most of the co-authorship networks, meaning that these networks are somehow already optimized with respect to knowledge spreading. However, as already mentioned, these results are still preliminary and need to be integrated with some objective performance measure of the collaboration networks under examination.

Although we limit its validation to the domain of R&D networks in this dissertation, we argue that our modeling framework is flexible and extendable to other collaboration networks, whose nodes can be unequivocally positioned in a knowledge space. Provided that a clear methodology to locate the nodes in a knowledge space and an appropriate performance measure of the network are unequivocally defined, our modeling framework offers the most complete and straightforward interpretation of the effects of knowledge exchange in a dynamically evolving collaboration network.

7.3 Future research

Even though our study answers all research questions that we have posed in Chapter 1, it inevitably leaves us with some other open questions. From a conceptual point of view, the first possible extension to our study would be to endogenize the rise-and-fall trend observed in R&D networks into our models.

As argued in Gulati *et al.* (2012), the rise and fall of R&D networks could be the sheer outcome of a knowledge recombination process, associated with the embeddedness into a network. Indeed, the possibility of knowledge recombination fuels the growth of the network, either by combining heterogeneous knowledge bases (e.g. Cowan and Jonard, 2004) or by granting access to multiple paths through which knowledge can reach the firm (König *et al.*, 2011). The same process of knowledge recombination may however set the the premises for the subsequent breakdown of the network. This is because recombination brings homogeneity into knowledge bases, consequently reducing the incentive for knowledge exchange and thus for alliance formation (Cowan and Jonard, 2004; Gulati *et al.*, 2012). Likewise, in a large network, the number of additional paths to which a firm gets access with an alliance is higher if the alliance is created with a firm which is already part of its component. In a situation where alliances are costly, this reduces the incentives to maintain bridging ties, thus contributing to the fragmentation of the network into many disconnected components (see König *et al.*, 2011, for a model generating a similar dynamics).

Next, although our network formation model captures many features of empirical collaboration networks, it can be further improved to account for other real world observations. One of the limitations is assuming fixed node labels; this condition could be relaxed by introducing a label decay, representing the exit of a firm from its circle of influence. Such an extension might be useful especially when validating a dataset with a longer time extension. A second limitation is that the alliance partners chosen by the initiators have no power in accepting this invitation; such a realistic attachment rule could be included in the model, at the price of requiring more parameters. In addition, a linking preference towards partners of partners could be added to the model, to better reproduce the observed effects of triadic closure in collaboration networks. Finally, we could include the study of assortativity and nestedness in the networks generated by our model: again, at the expense of adding more parameters or requiring more computational power, it would be possible to reproduce these even more sophisticated network properties.

With respect to our general modeling framework, a first extension is represented by the addition of a preferred knowledge distance for the agents to initiate a new collaboration, in agreement with theoretical arguments (Cohen and Levinthal, 1990) and with our empirical findings. To reproduce reality even better, at the expense of requiring more parameters, we could incorporate additional drift, self-motion or distancing mechanisms for the agents in the knowledge space. However, our results have shown that an approach mechanisms alone, coupled with a complex network dynamics, is already capable of reproducing both the left and the right tail of the knowledge distance shifts as a consequence of collaborations.

Another valid extension would be represented by the merger of our two-step validation procedure into a broader, more complex one-step procedure. This means that the network formation parameters and the knowledge exchange parameters would be estimated at the same time, by imposing both the network topology and the knowledge positions of the nodes. Such a procedure is formally more correct, because it would fully take into account the interdependencies between the network topology and the knowledge positions. On the other hand, varying all the parameters at the same time increases the dimensionality of the parameter space and requires more computational power.

As a logical consequence, this means that the study of the collaboration network *performance* would be more complete, as it would be expressed as a function of not only the knowledge exchange parameters, but also the topological network parameters. Moreover, the definition of the performance itself could be improved, or changed, by including other network-related measures that quantify the speed and/or efficiency of diffusion mechanisms on the network.

The definition of a comprehensive performance indicator, that takes into account the network *connectedness*, has several implications for quantifying the vulnerability and the systemic risk to which such systems are subject. As mentioned in Chapter 1, the interdependence between the network performance and the microscopic parameters is in most cases not trivial. A small change in the linking probabilities, for instance, can lead the system to a state that experiences a sudden drop in network connectedness, i.e. a state in which the removal of one single node can possibly cause a network breakdown.

Our work paves the way to a series of studies on performance, vulnerability and systemic risk in collaboration networks, which are necessary to understand the conditions leading to the systems' performance, as a function of both knowledge-related and network-related microscopic parameters.

Appendix A

Supplementary material to Chapter 2

Modularity in the pooled R&D network We report in Table A.1 the relative connectivities c_{ij} between the 18 largest industrial sectors that we have analyzed in our study. The pooled R&D network in the year 1995 has been used to compute these values.

The quantity c_{ij} indicates the ratio between the observed and the expected fraction of alliances connecting a firm in sector i to a firm in sector j. Values of c_{ij} greater than 1 suggest that the alliance probability between a firm in sector i and a firm in sector j is higher than one would expect with a random partner choice. On the contrary, when c_{ij} is smaller than 1, a firm in sector i forms alliances with firms in sector j with a smaller probability than a random partner choice.

	283	737	873	367	357	384	366	679	481	822	382	371	874	281	372	871	131	504
Pharmaceuticals (283)	**6.5**	0.1	3.1	0.1	0.1	2.2	0.2	1.0	0.1	3.6	1.6	0.2	0.4	2.4	0.0	0.0	0.0	0.8
Computer Software (737)	0.1	**3.4**	0.5	1.3	2.2	0.2	1.4	1.4	1.5	0.6	1.3	0.5	3.0	0.6	0.6	1.0	0.1	2.2
R&D, Lab and Testing (873)	3.1	0.5	**7.7**	0.3	0.2	2.1	0.2	2.1	0.9	4.0	2.8	0.9	4.3	2.3	0.0	2.3	0.0	1.7
Electronic Components (367)	0.1	1.3	0.3	**5.0**	3.2	1.0	3.3	1.7	0.8	0.7	1.3	0.0	0.8	0.6	0.0	3.3	0.0	1.8
Computer Hardware (357)	0.1	2.2	0.2	3.2	**3.5**	0.4	2.2	0.9	1.2	0.4	0.5	0.5	1.1	0.4	0.2	0.4	0.0	3.0
Medical Supplies (384)	2.2	0.2	2.1	1.0	0.4	**29.4**	0.2	1.8	0.0	1.3	2.7	0.8	0.0	0.0	0.0	0.0	0.0	0.0
Communications Equipment (366)	0.2	1.4	0.2	3.3	2.2	0.2	**7.2**	1.4	2.4	0.2	2.1	0.4	0.0	0.0	0.9	1.8	0.0	2.0
Investment Companies (679)	1.0	1.4	2.1	1.7	0.9	1.8	1.4	**9.1**	1.0	1.1	2.1	2.4	2.2	3.1	0.8	6.2	3.8	0.0
Telephone Communications (481)	0.1	1.5	0.9	0.8	1.2	0.0	2.4	1.0	**14.9**	0.8	1.0	0.4	4.7	0.0	0.6	0.0	2.0	0.0
Universities (822)	3.6	0.6	4.0	0.7	0.4	1.3	0.2	1.1	0.8	**6.9**	3.5	4.4	0.0	1.9	1.9	0.0	3.5	1.1
Laboratory Apparatus (382)	1.6	1.3	2.8	1.3	0.5	2.7	2.1	2.1	1.0	3.5	**10.7**	0.0	0.0	3.9	7.8	0.0	0.0	0.0
Motor Vehicles (371)	0.2	0.5	0.9	0.0	0.5	0.8	0.4	2.4	0.4	4.4	0.0	**60.8**	0.0	0.0	4.9	13.1	2.0	0.0
Management, Consulting, PR (874)	0.4	3.0	4.3	0.8	1.1	0.0	0.0	2.2	4.7	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0
Inorganic Chemicals (281)	2.4	0.6	2.3	0.6	0.4	0.0	0.0	3.1	0.0	1.9	3.9	0.0	0.0	**85.3**	4.3	0.0	5.1	0.0
Aircrafts and parts (372)	0.0	0.6	0.0	0.0	0.2	0.0	0.9	0.8	0.6	1.9	7.8	4.9	0.0	4.3	**76.7**	25.6	2.6	0.0
Engineer.,Architec.,Survey (871)	0.0	1.0	2.3	3.3	0.4	0.0	1.8	6.2	0.0	0.0	0.0	13.1	0.0	0.0	25.6	**0.0**	5.1	0.0
Crude Oil and Gas (131)	0.0	0.1	0.0	0.0	0.0	0.0	0.0	3.8	2.0	3.5	0.0	2.0	0.0	5.1	2.6	5.1	**118.1**	0.0
Profess. Equipment Wholesale (504)	0.8	2.2	1.7	1.8	3.0	0.0	2.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**10.4**

Table A.1: Relative connectivities between the 18 largest sectors.

As reported in Table A.1, most of the connectivity values in the main diagonal c_{ii}, that we define as the *intra-sector connectivities*, are greater than 1. On the contrary, the elements c_{ij} outside the main diagonal, that we define as the *inter-sector connectivities*, are instead smaller than 1, or anyway smaller than the corresponding diagonal element. In other words, most of the alliances we have tracked in the pooled R&D network occur within sectors rather than between different sectors. There are only two exceptions: Management-Consulting-PR and Engineering-Architecture-Survey, that – being service sectors – have a natural bias towards inter-sectoral alliances, rather than intra-sectoral alliances.

It is important to stress that we have not performed this analysis in order to maximize the modularity coefficient or any other indicator. Our aim was to assess the goodness of one possible partition of the R&D network (in this case, using the industrial sectors as communities) to study the evolution of the modularity coefficient computed on top of this partition over time. However, such cross-sector investigation sheds light on the alliance activity of the firms. Most of the alliances, being intra-sectoral, are characterized by homophily. But the existence of inter-sectoral alliances is further evidence that some firms serve as bridges between two or more different sectors, as we have already shown in Section 2.3.3 with respect to the assortativity.

Additional results for our econometric model The following tables report the complete results for our econometric model on the seven representative sectoral R&D networks that we have analyzed in Chapter 2, i.e. Pharmaceuticals (Table A.2), Computer Hardware (Table A.3), Communications Equipment (Table A.4), Electronic Components (Table A.5), Medical Supplies (Table A.6), Computer Software (Table A.7) and R&D, Laboratory and Testing (Table A.8).

	A	B	C	AB	AC	BC	ABC
(Intercept)	$\mathbf{-4.966}^{***}$	$\mathbf{-6.271}^{***}$	$\mathbf{-4.781}^{***}$	$\mathbf{-5.347}^{***}$	$\mathbf{-4.060}^{***}$	$\mathbf{-5.240}^{***}$	$\mathbf{-4.467}^{***}$
	(0.192)	(0.183)	(0.291)	(0.206)	(0.301)	(0.323)	(0.341)
newlinks	$\mathbf{0.508}^{***}$	$\mathbf{0.472}^{***}$	$\mathbf{0.494}^{***}$	$\mathbf{0.483}^{***}$	$\mathbf{0.508}^{***}$	$\mathbf{0.471}^{***}$	$\mathbf{0.481}^{***}$
	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)	(0.010)
same_nation	$\mathbf{0.224}^{***}$			0.163^{*}	$\mathbf{0.222}^{**}$		0.169^{*}
	(0.068)			(0.068)	(0.068)		(0.068)
same_sic	$\mathbf{0.371}^{***}$			$\mathbf{0.448}^{***}$	$\mathbf{0.383}^{***}$		$\mathbf{0.454}^{***}$
	(0.079)			(0.079)	(0.079)		(0.079)
past_alliances	$\mathbf{0.273}^{***}$			0.139	$\mathbf{0.263}^{***}$		0.128
	(0.061)			(0.098)	(0.063)		(0.101)
tech_distance	$\mathbf{-2.668}^{***}$			$\mathbf{-2.103}^{***}$	$\mathbf{-2.618}^{***}$		$\mathbf{-2.095}^{***}$
	(0.137)			(0.139)	(0.138)		(0.139)
inverse_shortest_pl		$\mathbf{2.920}^{***}$		$\mathbf{2.189}^{***}$		$\mathbf{2.910}^{***}$	$\mathbf{2.218}^{***}$
		(0.127)		(0.143)		(0.132)	(0.146)
closeness_aritm_mean		$\mathbf{-0.507}^{***}$		$\mathbf{-0.443}^{**}$		$\mathbf{-0.684}^{***}$	$\mathbf{-0.674}^{***}$
		(0.151)		(0.156)		(0.171)	(0.179)
closeness_difference		$\mathbf{0.604}^{***}$		$\mathbf{0.535}^{***}$		$\mathbf{0.684}^{***}$	$\mathbf{0.549}^{***}$
		(0.100)		(0.105)		(0.163)	(0.162)
delta_closeness			$\mathbf{-0.860}^{***}$		$\mathbf{-0.533}^{**}$	$\mathbf{-0.622}^{**}$	-0.435^{*}
			(0.162)		(0.165)	(0.203)	(0.204)
delta_eigenvalue			0.026		0.013	0.050^{*}	0.036
			(0.019)		(0.019)	(0.022)	(0.022)
delta_harmonic_aspl			0.051		-0.083	-0.092	-0.200^{*}
			(0.082)		(0.082)	(0.099)	(0.095)
AIC	9166.072	9336.770	9744.345	8941.128	9153.738	9329.115	8935.185
BIC	9432.478	9592.519	10000.095	9239.502	9452.113	9616.833	9265.528
Log Likelihood	-4558.036	-4644.385	-4848.173	-4442.564	-4548.869	-4637.557	-4436.593
Deviance	9116.072	9288.770	9696.345	8885.128	9097.738	9275.115	8873.185
Num. obs.	313709	313709	313709	313709	313709	313709	313709

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table A.2: Econometric model ABC (including all variable groups) for the Pharmaceuticals sectoral R&D network. The coefficients with p-value smaller than 0.01 are reported in bold character.

	A	B	C	AB	AC	BC	ABC
(Intercept)	**−5.102*****	**−5.712*****	**−4.284*****	**−5.345*****	**−4.118*****	**−4.854*****	**−4.588*****
	(0.190)	(0.184)	(0.205)	(0.197)	(0.211)	(0.225)	(0.234)
newlinks	**0.219*****	**0.199*****	**0.223*****	**0.197*****	**0.216*****	**0.200*****	**0.199*****
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
same_nation	**0.675*****			**0.651*****	**0.702*****		**0.648*****
	(0.056)			(0.056)	(0.056)		(0.056)
same_sic	**0.365*****			**0.332*****	**0.361*****		**0.329*****
	(0.067)			(0.067)	(0.067)		(0.067)
past_alliances	**0.208*****			**0.111****	**0.179*****		**0.116****
	(0.032)			(0.037)	(0.033)		(0.037)
tech_distance	**−1.585*****			**−1.334*****	**−1.449*****		**−1.298*****
	(0.109)			(0.110)	(0.109)		(0.110)
inverse_shortest_pl		**1.930*****		**1.519*****		**1.981*****	**1.569*****
		(0.126)		(0.128)		(0.124)	(0.127)
closeness_aritm_mean		0.179$^{\cdot}$		0.179*		−0.047	−0.042
		(0.093)		(0.090)		(0.101)	(0.099)
closeness_difference		**0.219*****		**0.162****		0.291*	0.246$^{\cdot}$
		(0.056)		(0.055)		(0.131)	(0.130)
delta_closeness			**−1.228*****		**−1.061*****	**−0.775*****	**−0.733*****
			(0.144)		(0.140)	(0.193)	(0.189)
delta_eigenvalue			**0.118*****		**0.105*****	**0.117*****	**0.105*****
			(0.023)		(0.023)	(0.022)	(0.022)
delta_harmonic_aspl			**−1.410*****		**−1.229*****	**−1.295*****	**−1.170*****
			(0.227)		(0.220)	(0.222)	(0.220)
AIC	11497.060	11613.424	11809.827	11248.290	11348.919	11538.025	11184.447
BIC	11749.789	11856.044	12052.448	11531.347	11631.976	11810.973	11497.831
Log Likelihood	-5723.530	-5782.712	-5880.914	-5596.145	-5646.459	-5742.013	-5561.223
Deviance	11447.060	11565.424	11761.827	11192.290	11292.919	11484.025	11122.447
Num. obs.	181529	181529	181529	181529	181529	181529	181529

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table A.3: Econometric model ABC (including all variable groups) for the Computer Hardware sectoral R&D network. The coefficients with p-value smaller than 0.01 are reported in bold character.

	A	B	C	AB	AC	BC	ABC
(Intercept)	**−5.078*****	**−5.561*****	**−3.463*****	**−5.431*****	**−3.504*****	**−4.694*****	**−4.636*****
	(0.199)	(0.206)	(0.223)	(0.220)	(0.227)	(0.273)	(0.281)
newlinks	**0.259*****	**0.238*****	**0.260*****	**0.237*****	**0.254*****	**0.237*****	**0.237*****
	(0.005)	(0.006)	(0.005)	(0.006)	(0.005)	(0.006)	(0.006)
same_nation	**0.736*****			**0.715*****	**0.744*****		**0.695*****
	(0.059)			(0.059)	(0.059)		(0.059)
same_sic	**0.512*****			**0.464*****	**0.497*****		**0.448*****
	(0.074)			(0.075)	(0.074)		(0.075)
past_alliances	0.047			−0.143*	−0.009		−0.157*
	(0.061)			(0.069)	(0.063)		(0.070)
tech_distance	**−1.386*****			**−1.067*****	**−1.244*****		**−1.077*****
	(0.120)			(0.122)	(0.120)		(0.122)
inverse_shortest_pl		**1.442*****		**1.130*****		**1.626*****	**1.315*****
		(0.126)		(0.130)		(0.129)	(0.133)
closeness_aritm_mean		**0.340*****		**0.345*****		**0.192****	**0.203****
		(0.071)		(0.069)		(0.074)	(0.073)
closeness_difference		0.000		−0.021		−0.114	−0.130
		(0.050)		(0.051)		(0.104)	(0.105)
delta_closeness			**−0.904*****		**−0.827*****	**−0.350****	**−0.329****
			(0.094)		(0.091)	(0.125)	(0.125)
delta_eigenvalue			**0.090*****		**0.091*****	**0.139*****	**0.132*****
			(0.019)		(0.019)	(0.019)	(0.019)
delta_harmonic_aspl			**−1.049*****		**−0.939*****	**−0.825*****	**−0.747*****
			(0.185)		(0.181)	(0.182)	(0.179)
AIC	9345.886	9412.960	9537.443	9124.595	9190.386	9341.452	9061.810
BIC	9587.563	9644.970	9769.453	9395.274	9461.064	9602.463	9361.489
Log Likelihood	-4647.943	-4682.480	-4744.722	-4534.298	-4567.193	-4643.726	-4499.905
Deviance	9295.886	9364.960	9489.443	9068.595	9134.386	9287.452	8999.810
Num. obs.	116668	116668	116668	116668	116668	116668	116668

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table A.4: Econometric model ABC (including all variable groups) for the Communications Equipment sectoral R&D network. The coefficients with p-value smaller than 0.01 are reported in bold character.

	A	B	C	AB	AC	BC	ABC
(Intercept)	**−5.266*****	**−5.951*****	**−3.801*****	**−5.696*****	**−3.697*****	**−4.995*****	**−4.852*****
	(0.181)	(0.197)	(0.231)	(0.210)	(0.236)	(0.281)	(0.288)
newlinks	**0.238*****	**0.213*****	**0.238*****	**0.212*****	**0.231*****	**0.212*****	**0.212*****
	(0.005)	(0.005)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
same_nation	**0.804*****			**0.767*****	**0.815*****		**0.761*****
	(0.054)			(0.055)	(0.055)		(0.055)
same_sic	**0.375*****			**0.354*****	**0.380*****		**0.348*****
	(0.067)			(0.068)	(0.067)		(0.068)
past_alliances	**0.175*****			−0.002	0.123*		−0.020
	(0.048)			(0.062)	(0.051)		(0.064)
tech_distance	**−1.534*****			**−1.255*****	**−1.413*****		**−1.234*****
	(0.119)			(0.120)	(0.119)		(0.120)
inverse_shortest_pl		**1.710*****		**1.354*****		**1.877*****	**1.523*****
		(0.119)		(0.123)		(0.121)	(0.125)
closeness_aritm_mean		**0.532*****		**0.542*****		0.300*	**0.333****
		(0.114)		(0.111)		(0.125)	(0.123)
closeness_difference		0.138$^{\cdot}$		0.070		−0.083	−0.157
		(0.073)		(0.073)		(0.146)	(0.146)
delta_closeness			**−1.134*****		**−1.052*****	−0.384*	−0.332$^{\cdot}$
			(0.149)		(0.146)	(0.183)	(0.181)
delta_eigenvalue			**0.054****		**0.056****	**0.109*****	**0.102*****
			(0.020)		(0.020)	(0.019)	(0.019)
delta_harmonic_aspl			**−1.004*****		**−0.905*****	**−0.899*****	**−0.813*****
			(0.161)		(0.159)	(0.167)	(0.164)
AIC	11644.651	11715.646	11934.436	11354.304	11488.450	11656.961	11304.548
BIC	11898.437	11959.280	12178.071	11638.544	11772.690	11931.050	11619.243
Log Likelihood	-5797.325	-5833.823	-5943.218	-5649.152	-5716.225	-5801.481	-5621.274
Deviance	11594.651	11667.646	11886.436	11298.304	11432.450	11602.961	11242.548
Num. obs.	189365	189365	189365	189365	189365	189365	189365

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table A.5: Econometric model ABC (including all variable groups) for the Electronic Components sectoral R&D network. The coefficients with p-value smaller than 0.01 are reported in bold character.

	A	B	C	AB	AC	BC	ABC
(Intercept)	**−4.380*****	**−6.262*****	**−2.416****	**−5.155*****	**−3.500*****	**−3.978****	−2.634*
	(0.769)	(0.735)	(0.836)	(0.799)	(0.863)	(1.254)	(1.289)
newlinks	**0.875*****	**0.971*****	**0.963*****	**0.930*****	**0.969****	**0.989*****	**0.941*****
	(0.075)	(0.076)	(0.077)	(0.075)	(0.074)	(0.079)	(0.077)
same_nation	−0.102			−0.016	0.050		−0.004
	(0.241)			(0.242)	(0.414)		(0.250)
same_sic	0.595*			**0.796****	**0.375****		**0.752****
	(0.279)			(0.279)	(0.146)		(0.282)
past_alliances	0.371*			0.232	0.295*		0.222
	(0.174)			(0.255)	(0.150)		(0.249)
tech_distance	**−3.326*****			**−2.636*****	**−2.812*****		**−2.721*****
	(0.529)			(0.532)	(0.551)		(0.543)
inverse_shortest_pl		**3.071*****		**2.474*****		**2.703*****	**2.089*****
		(0.348)		(0.362)		(0.382)	(0.393)
closeness_aritm_mean		0.013		0.016		−0.031	−0.031
		(0.017)		(0.017)		(0.029)	(0.027)
closeness_difference		0.007		0.012		0.027	0.032
		(0.014)		(0.014)		(0.021)	(0.021)
delta_closeness			**−0.149*****		**−0.146*****	−0.094*	−0.099*
			(0.025)		(0.022)	(0.046)	(0.045)
delta_eigenvalue			−0.116		−0.113	−0.131	−0.088
			(0.085)		(0.079)	(0.100)	(0.102)
delta_harmonic_aspl			0.009		0.008	−0.024	−0.051
			(0.198)		(0.310)	(0.227)	(0.234)
AIC	655.193	663.227	695.624	609.758	630.024	658.479	605.090
BIC	838.657	839.353	871.749	815.237	835.504	856.620	832.585
Log Likelihood	-302.597	-307.614	-323.812	-276.879	-307.012	-302.239	-271.545
Deviance	605.193	615.227	647.624	553.758	647.024	604.479	543.090
Num. obs.	11368	11368	11368	11368	11368	11368	11368

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table A.6: Econometric model ABC (including all variable groups) for the Medical Supplies sectoral R&D network. The coefficients with p-value smaller than 0.01 are reported in bold character.

	A	B	C	AB	AC	BC	ABC
(Intercept)	**−4.971*****	**−5.939*****	**−4.598*****	**−5.316*****	**−4.102*****	**−5.229*****	**−4.647*****
	(0.192)	(0.189)	(0.206)	(0.200)	(0.212)	(0.225)	(0.234)
newlinks	**0.196*****	**0.178*****	**0.206*****	**0.173*****	**0.194*****	**0.180*****	**0.175*****
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
same_nation	**0.517*****			**0.506*****	**0.544*****		**0.496*****
	(0.055)			(0.056)	(0.055)		(0.056)
same_sic	**0.401*****			**0.361*****	**0.399*****		**0.349*****
	(0.067)			(0.067)	(0.067)		(0.067)
past_alliances	−0.112			−0.252*	−0.156		−0.241*
	(0.099)			(0.111)	(0.103)		(0.112)
tech_distance	**−1.893*****			**−1.652*****	**−1.807*****		**−1.630*****
	(0.105)			(0.106)	(0.105)		(0.106)
inverse_shortest_pl		**1.926*****		**1.592*****		**2.188*****	**1.833*****
		(0.118)		(0.117)		(0.119)	(0.120)
closeness_aritm_mean		**0.538*****		**0.503*****		0.189	0.149
		(0.120)		(0.116)		(0.123)	(0.122)
closeness_difference		**0.338*****		**0.295*****		0.425*	**0.439****
		(0.073)		(0.072)		(0.170)	(0.170)
delta_closeness			**−1.232*****		**−1.073*****	**−0.843*****	**−0.847*****
			(0.168)		(0.162)	(0.239)	(0.240)
delta_eigenvalue			**0.074*****		**0.064*****	**0.123*****	**0.109*****
			(0.019)		(0.019)	(0.016)	(0.016)
delta_harmonic_aspl			**−1.580*****		**−1.415*****	**−1.137*****	**−1.047*****
			(0.237)		(0.232)	(0.240)	(0.239)
AIC	12297.023	12425.920	12721.241	11968.114	12169.407	12342.446	11900.390
BIC	12554.659	12673.250	12968.572	12256.666	12457.959	12620.693	12219.859
Log Likelihood	-6123.511	-6188.960	-6336.620	-5956.057	-6056.703	-6144.223	-5919.195
Deviance	12247.023	12377.920	12673.241	11912.114	12113.407	12288.446	11838.390
Num. obs.	220896	220896	220896	220896	220896	220896	220896

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table A.7: Econometric model ABC (including all variable groups) for the Computer Software sectoral R&D network. The coefficients with p-value smaller than 0.01 are reported in bold character.

	A	B	C	AB	AC	BC	ABC
(Intercept)	**−4.496*****	**−6.123*****	**−3.890*****	**−4.825*****	**−3.332*****	**−5.489*****	**−4.537*****
	(0.179)	(0.185)	(0.292)	(0.202)	(0.296)	(0.415)	(0.410)
newlinks	**0.341*****	**0.309*****	**0.359*****	**0.309*****	**0.338*****	**0.309*****	**0.307*****
	(0.008)	(0.009)	(0.008)	(0.009)	(0.009)	(0.009)	(0.009)
same_nation	**0.236*****			**0.206****	**0.225*****		**0.206****
	(0.066)			(0.067)	(0.066)		(0.067)
same_sic	**0.718*****			**0.701*****	**0.734*****		**0.700*****
	(0.075)			(0.076)	(0.075)		(0.076)
past_alliances	0.123$^{\cdot}$			0.059	0.108		0.060
	(0.073)			(0.087)	(0.076)		(0.088)
tech_distance	**−2.989*****			**−2.578*****	**−2.903*****		**−2.576*****
	(0.135)			(0.139)	(0.136)		(0.139)
inverse_shortest_pl		**2.644*****		**1.633*****		**2.681*****	**1.721*****
		(0.123)		(0.134)		(0.126)	(0.137)
closeness_aritm_mean		0.165		0.309		0.197	0.326
		(0.242)		(0.238)		(0.330)	(0.321)
closeness_difference		**0.455****		0.281$^{\cdot}$		0.197	0.002
		(0.156)		(0.160)		(0.198)	(0.198)
delta_closeness			**−1.578*****		**−0.835*****	−0.431$^{\cdot}$	−0.098
			(0.217)		(0.206)	(0.241)	(0.234)
delta_eigenvalue			0.006		−0.013	**0.089*****	0.063*
			(0.022)		(0.022)	(0.025)	(0.025)
delta_harmonic_aspl			0.077		−0.099	−0.033	−0.139
			(0.087)		(0.087)	(0.093)	(0.095)
AIC	9536.689	10000.926	10392.194	9347.891	9503.648	9993.217	9344.910
BIC	9801.499	10255.143	10646.410	9644.477	9800.234	10279.211	9673.273
Log Likelihood	-4743.345	-4976.463	-5172.097	-4645.945	-4723.824	-4969.609	-4641.455
Deviance	9486.689	9952.926	10344.194	9291.891	9447.648	9939.217	9282.910
Num. obs.	294304	294304	294304	294304	294304	294304	294304

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\cdot}p < 0.1$

Table A.8: Econometric model ABC (including all variable groups) for the R&D, Lab and Testing sectoral R&D network. The coefficients with p-value smaller than 0.01 are reported in bold character.

Appendix B

Supplementary material to Chapter 3

Network visualization and emergence of communities We report here the visual network representations and the cluster size distributions for five sectoral R&D networks – Computer Hardware (Fig. B.1), Communications Equipment (Fig. B.2), Electronic Components (Fig. B.3), Computer Software (Fig. B.4) and R&D, Laboratory and Testing (Fig. B.5) – as well as five co-authorship networks – quantum mechanics, field theories and special relativity (Fig. B.6), general relativity and gravitation (Fig. B.7), optics (Fig. B.8), electronic transport in condensed matter (Fig. B.9) and superconductivity (Fig. B.10).

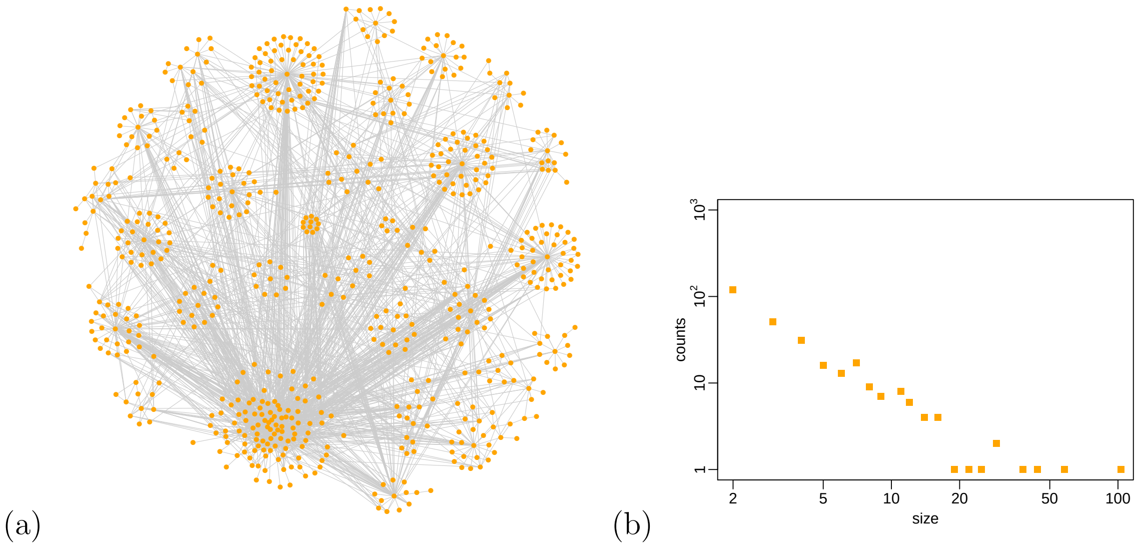


Figure B.1: (a) Visual representation of the computer hardware sectoral R&D network. (b) Size distribution of the network clusters.

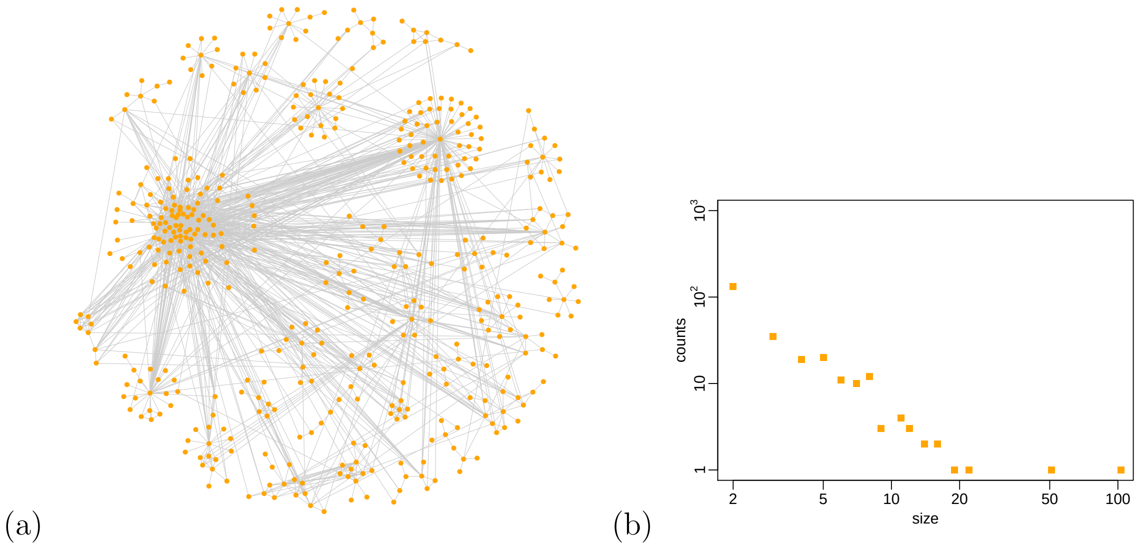


Figure B.2: (a) Visual representation of the communications equipment sectoral R&D network. (b) Size distribution of the network clusters.

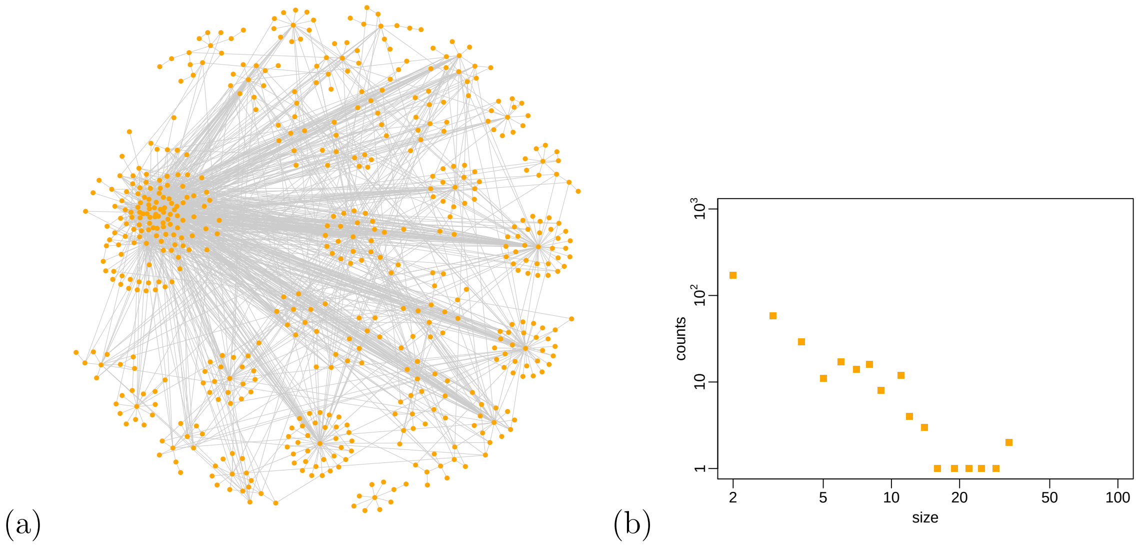


Figure B.3: (a) Visual representation of the electronic components sectoral R&D network. (b) Size distribution of the network clusters.

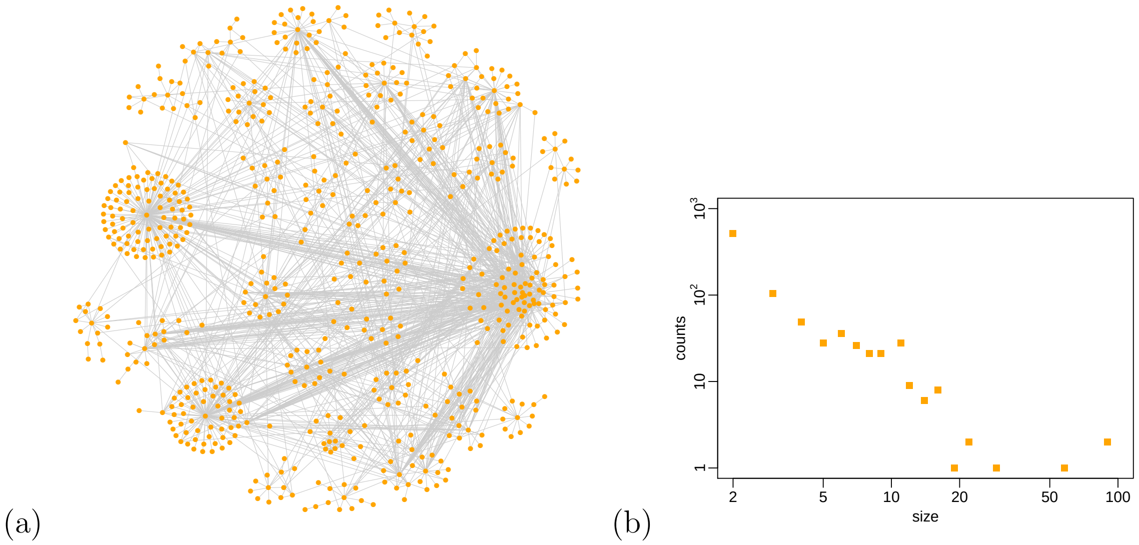


Figure B.4: (a) Visual representation of the computer software sectoral R&D network. (b) Size distribution of the network clusters.

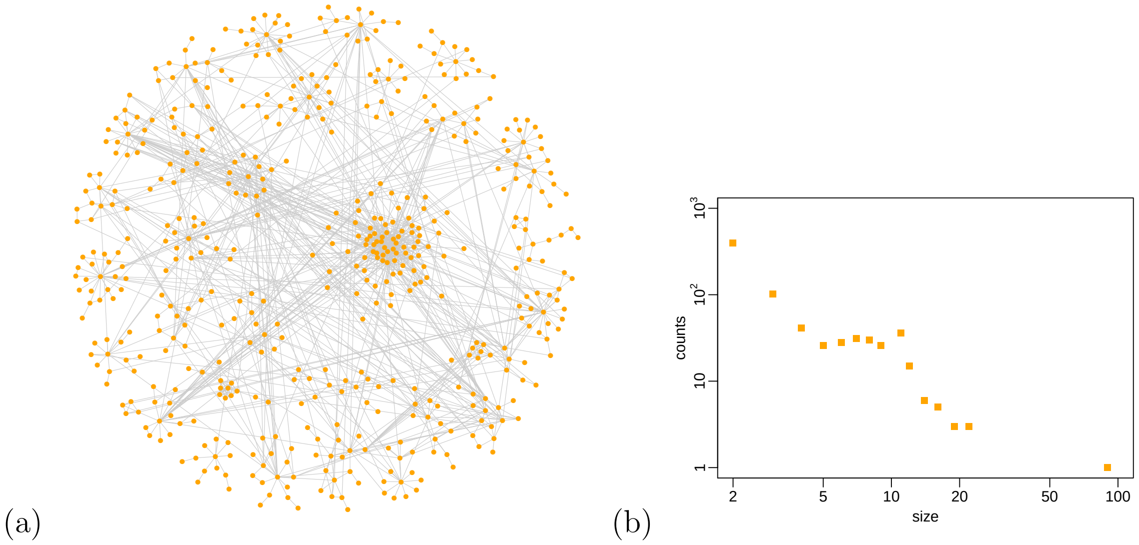


Figure B.5: (a) Visual representation of the R&D, laboratory & testing sectoral R&D network. (b) Size distribution of the network clusters.

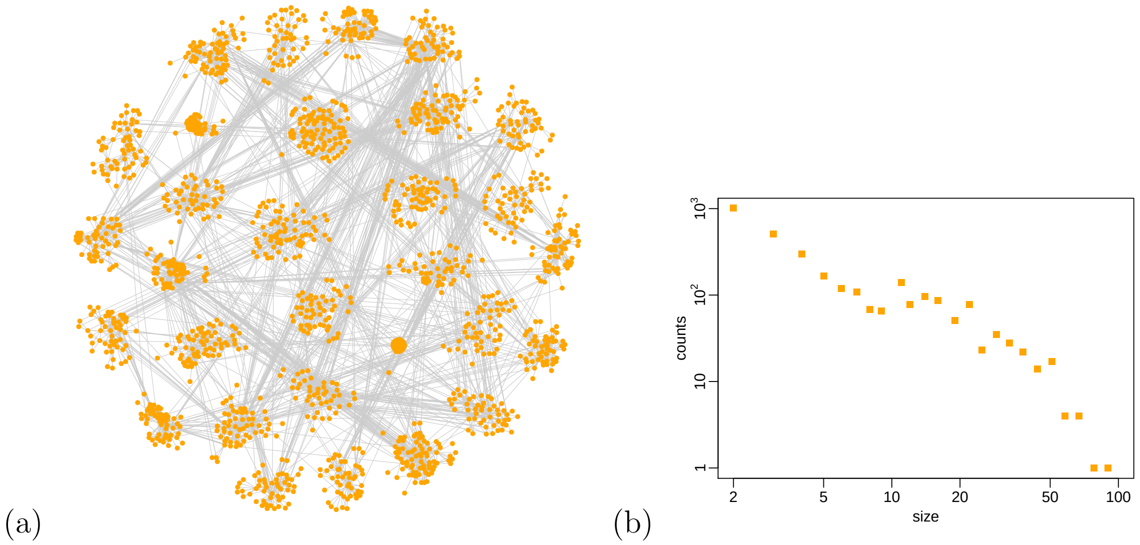


Figure B.6: (a) Visual representation of the co-authorship network in quantum mechanics, field theories and special relativity. (b) Size distribution of the network clusters.

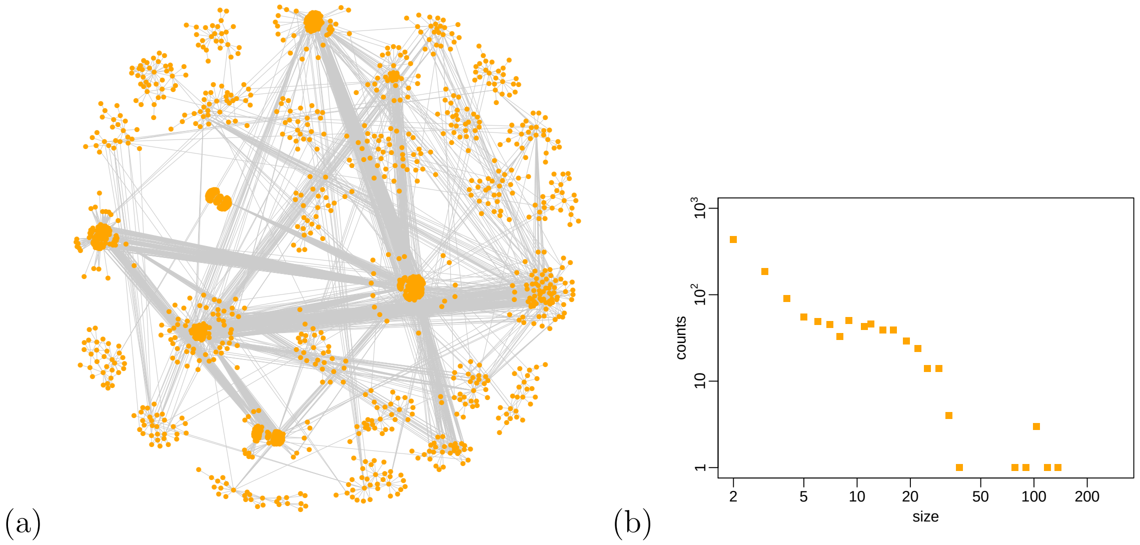


Figure B.7: (a) Visual representation of the co-authorship network in general relativity and gravitation. (b) Size distribution of the network clusters.

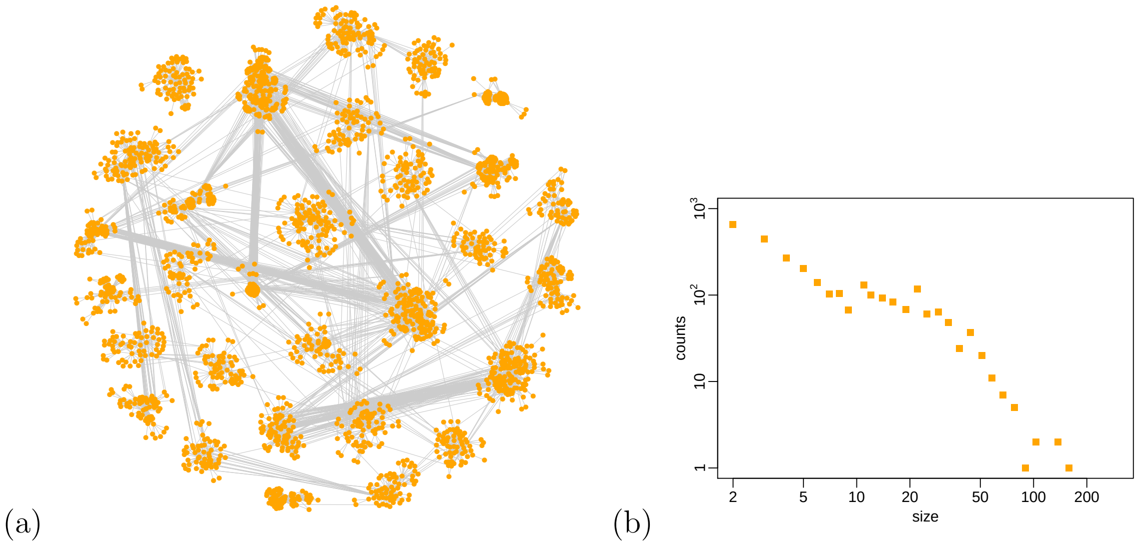


Figure B.8: (a) Visual representation of the co-authorship network in optics. (b) Size distribution of the network clusters.

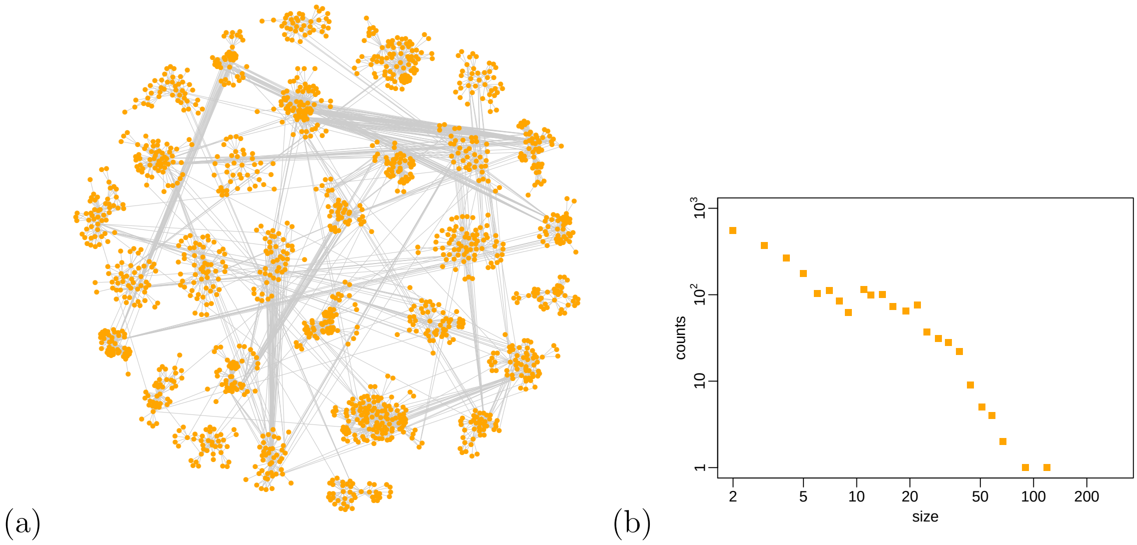


Figure B.9: (a) Visual representation of the co-authorship network in electronic transport in condensed matter. (b) Size distribution of the network clusters.

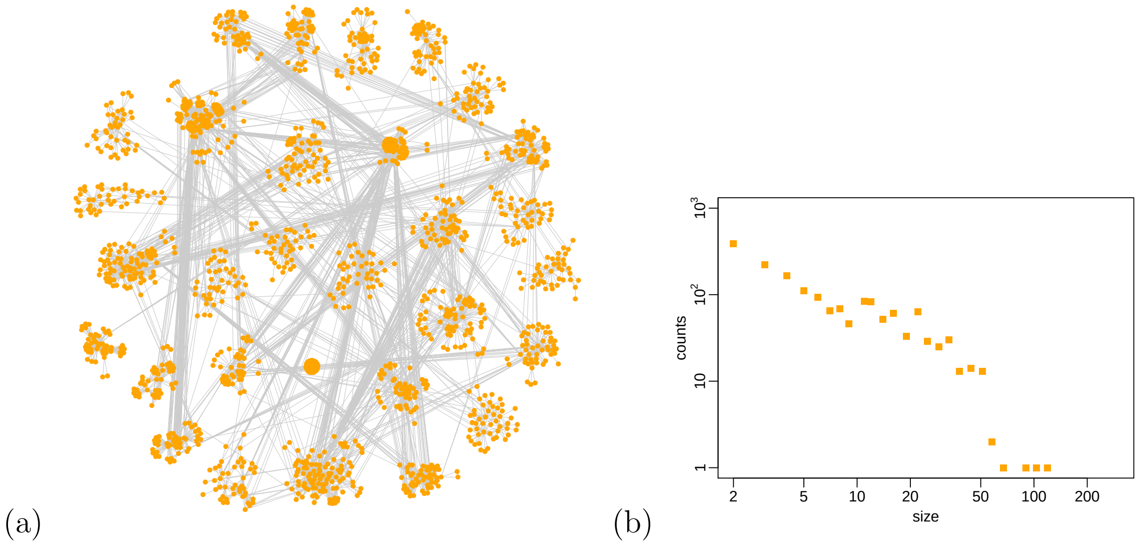


Figure B.10: (a) Visual representation of the co-authorship network in superconductivity. (b) Size distribution of the network clusters.

Appendix C

Supplementary material to Chapter 4

Numerical simulations results For each of the 684,000 computer simulations we run, we test the resulting generated R&D network with respect to three properties: average degree $\langle k \rangle$, average path length $\langle l \rangle$ and global clustering coefficient C. In Fig. C.1 we show how these three quantities are distributed across all the 684,000 realizations and we compare them with the observed values $\langle k \rangle^{OBS}$, $\langle l \rangle^{OBS}$ and C^{OBS}.

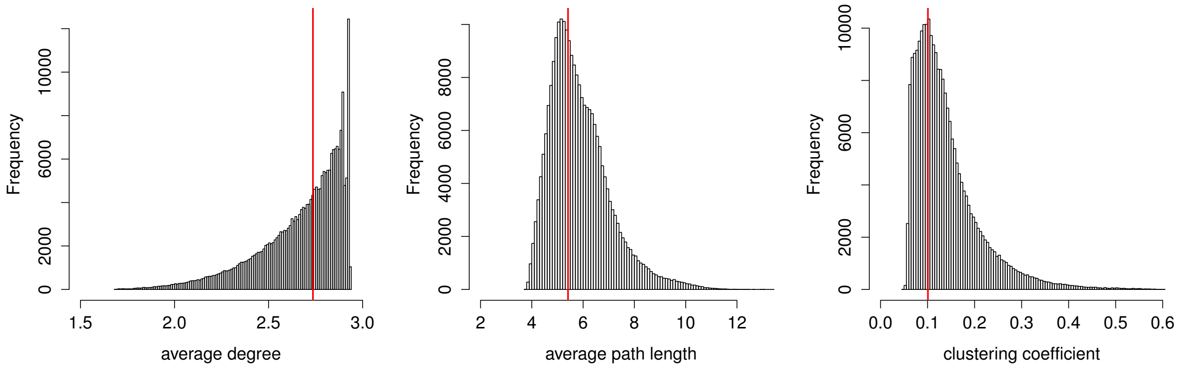


Figure C.1: Distributions of average degree $\langle k \rangle$, average path length $\langle l \rangle$ and global clustering coefficient C across all 684,000 runs of our model (each of the 3,420 points in the parameter space has been explored 200 times). The vertical red lines represent the observed values $\langle k \rangle^{OBS}$, $\langle l \rangle^{OBS}$ and C^{OBS} in the empirical R&D network.

We find that the global clustering coefficient and the average path length distributions are peaked around the observed values. However, the average degree distribution does not display any peak, despite being relatively narrow and centered around the real value (note the values on the x-axis in Fig. C.1). The fact that these three distributions are centered around the real values testifies that our model well captures the topology of the observed network for a large set of free parameters, despite we have imposed only a few features of the network (number of nodes N and alliances E, and the distributions of node activities a_i and partners per alliance m). At the same time, the distributions of $\langle k \rangle$, $\langle l \rangle$ and C are not excessively narrow, showing that the we can meaningfully perform an exploration – and consequently a fit – of the free parameters of our model.

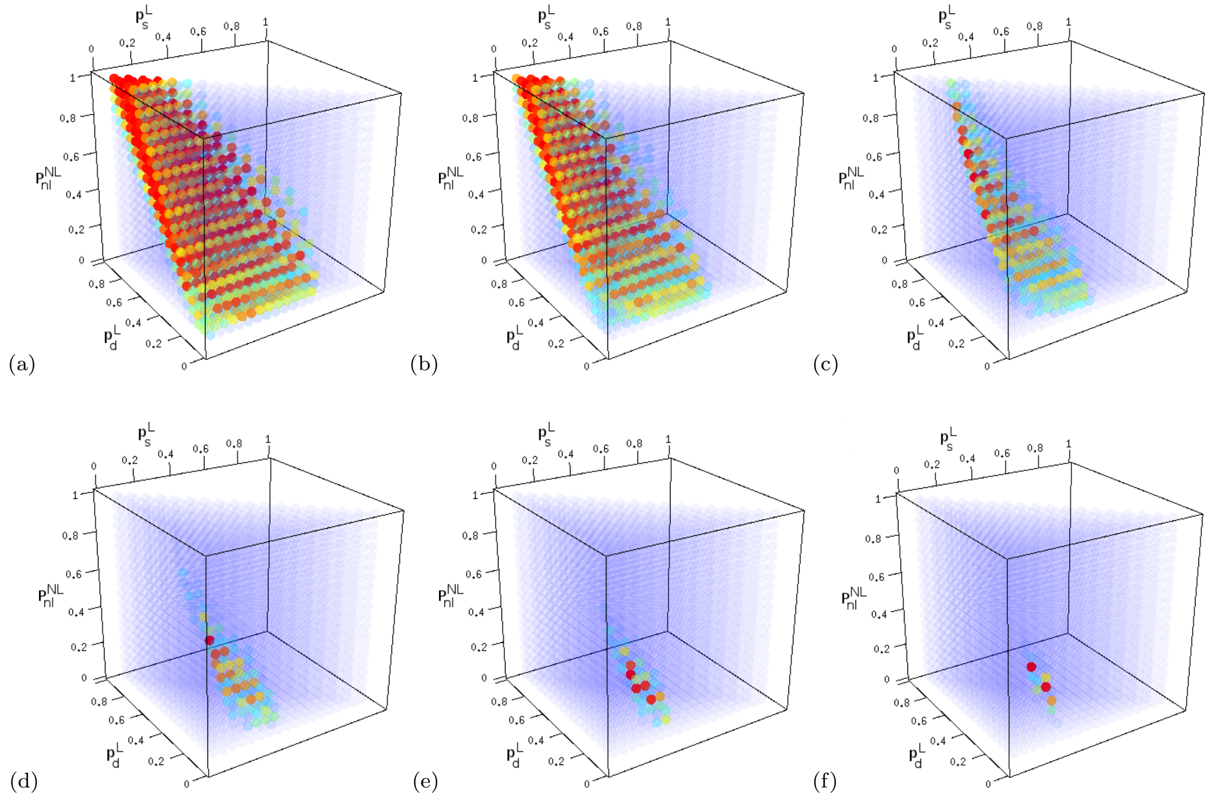


Figure C.2: Likelihood scores for all points in the parameter space, for ε^0 equal to 10% (a), 8% (b), 5% (c), 3% (d), 2% (e) and 1% (f).

The error threshold value ε^0 we impose for the computation of the Likelihood score influences the number of points in the parameter space that fulfill our matching criteria. Obviously, by decreasing ε^0, we observe a smaller number of points displaying high likelihood scores, as we could expect, because a better representation of reality is required. In Fig. C.2 we show the Likelihood scores of every point in the parameter space for six different values of ε^0, ranging from 1% to 10%. For our analysis, we take a conservative approach and fix $\varepsilon^0 = 2\%$.

Microscopic measures on all tested collaboration networks We report all the microscopic features that our agent-based mode is able to reproduce, for all the collaboration networks under examination.

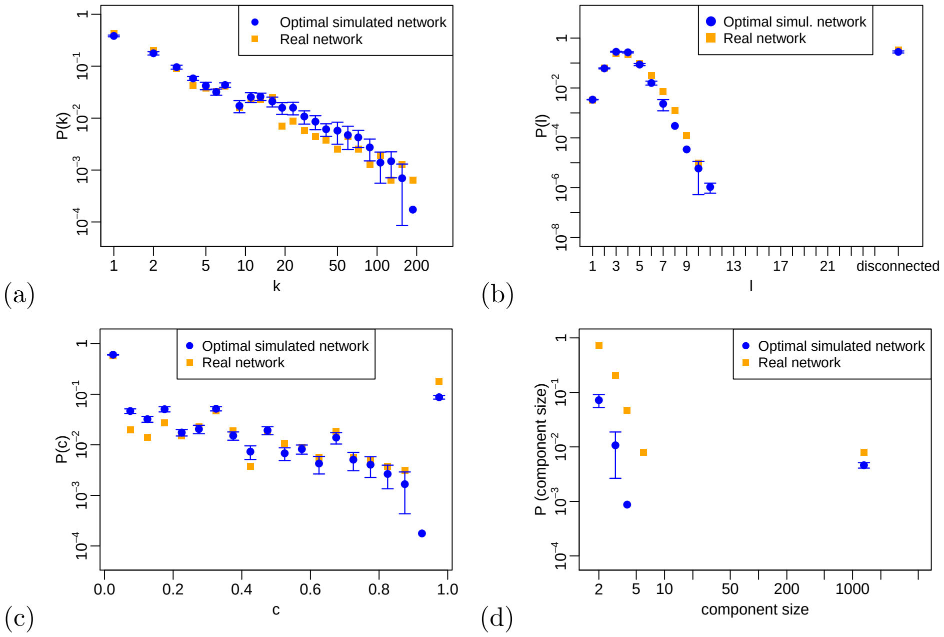


Figure C.3: Computer hardware R&D network (SIC code 357). Distributions of node degrees (a), path lengths (b), local clustering coefficients (c) and component sizes (d) for the real and the optimal simulated networks.

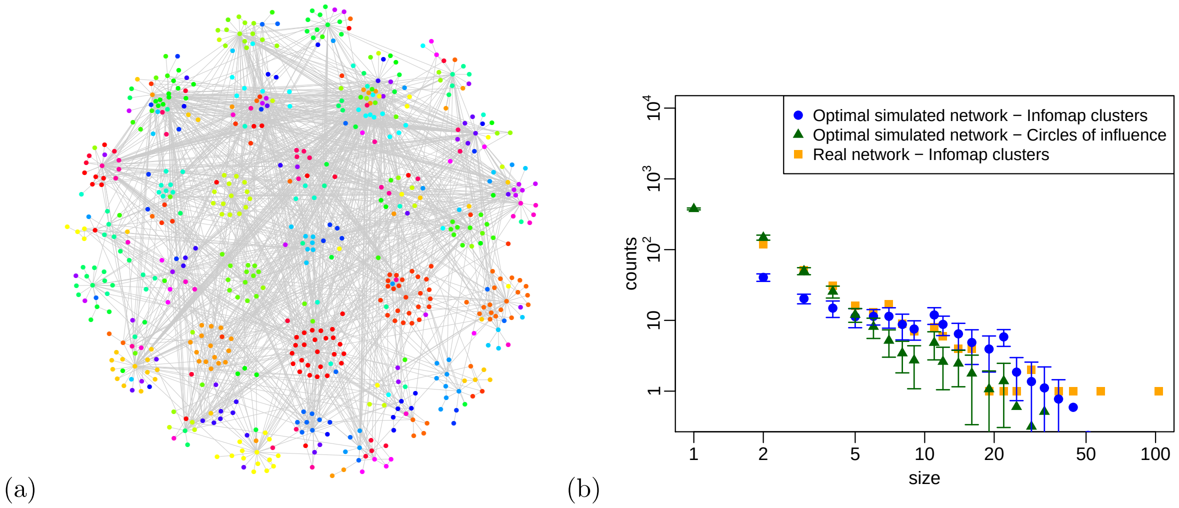


Figure C.4: Computer hardware R&D network (SIC code 357). (a) Visual representation of one realization of the optimal simulated network. (b) Size distribution of i. the circles of influence in all realizations of the optimal simulated network, ii. the Infomap clusters in all realizations of the optimal simulated network and iii. the Infomap clusters in the empirical network.

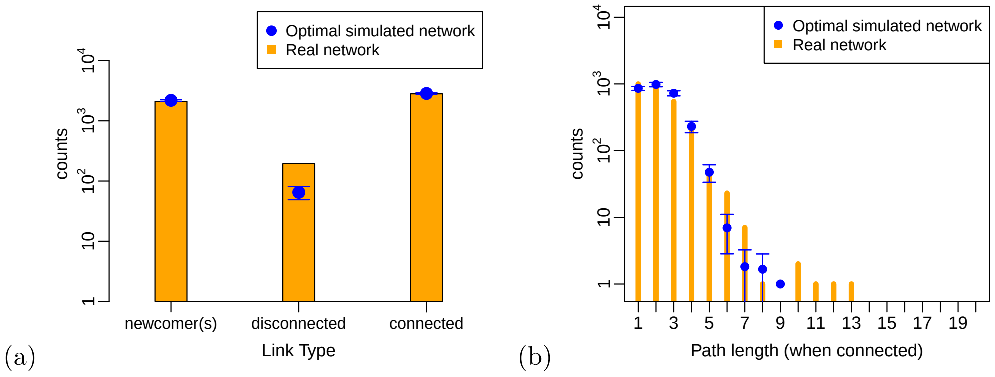


Figure C.5: Computer hardware R&D network (SIC code 357). (a) Distribution of link types for empirical and simulated networks. (b) Distribution of path lengths at the moment of link formation (only for nodes belonging to the same connected component).

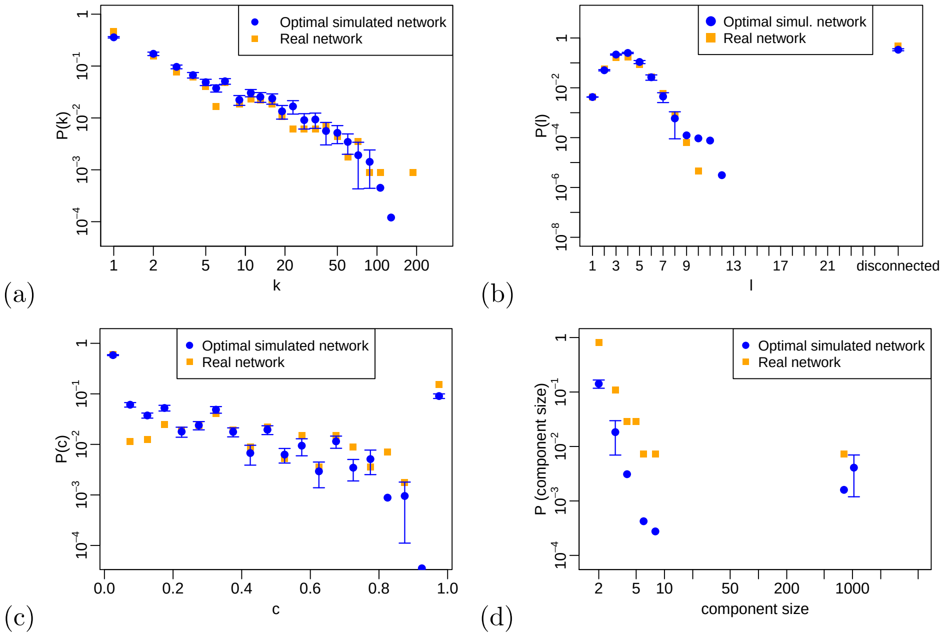


Figure C.6: Communications equipment R&D network (SIC code 366). Distributions of node degrees (a), path lengths (b), local clustering coefficients (c) and component sizes (d) for the real and the optimal simulated networks.

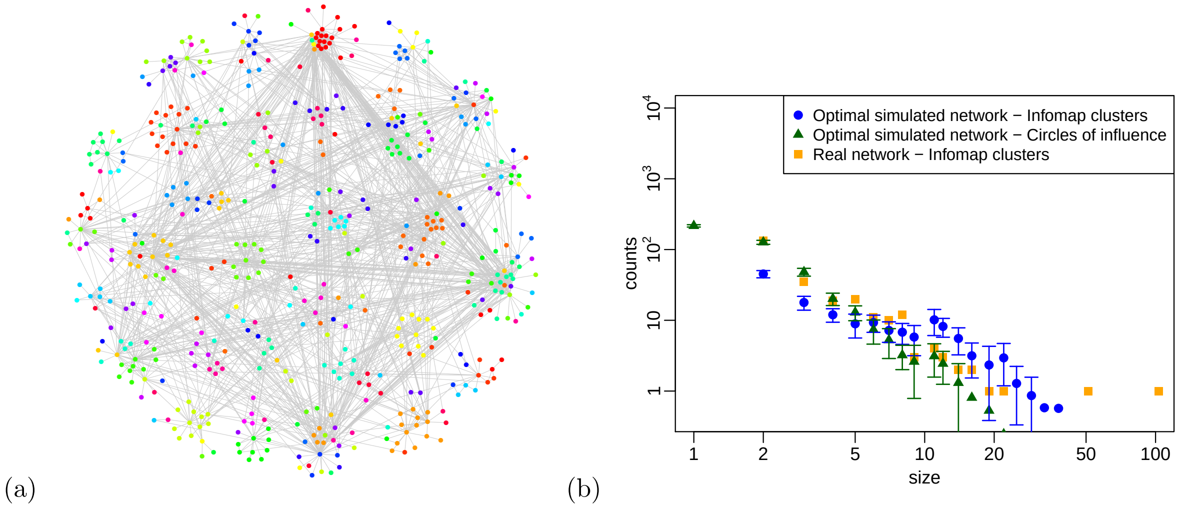


Figure C.7: Communications equipment R&D network (SIC code 366). (a) Visual representation of one realization of the optimal simulated network. (b) Size distribution of i. the circles of influence in all realizations of the optimal simulated network, ii. the Infomap clusters in all realizations of the optimal simulated network and iii. the Infomap clusters in the empirical network.

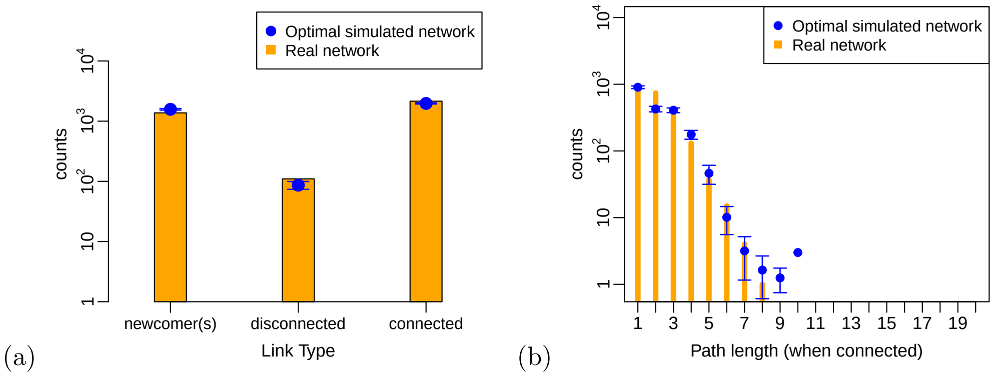


Figure C.8: Communications equipment R&D network (SIC code 366). (a) Distribution of link types for empirical and simulated networks. (b) Distribution of path lengths at the moment of link formation (only for nodes belonging to the same connected component).

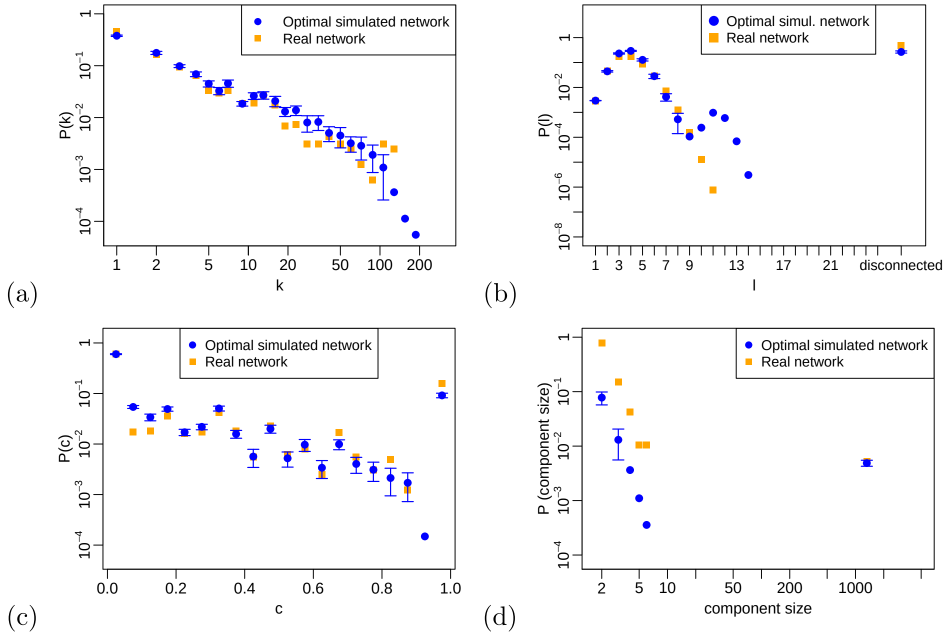


Figure C.9: Electronic components R&D network (SIC code 367). Distributions of node degrees (a), path lengths (b), local clustering coefficients (c) and component sizes (d) for the real and the optimal simulated networks.

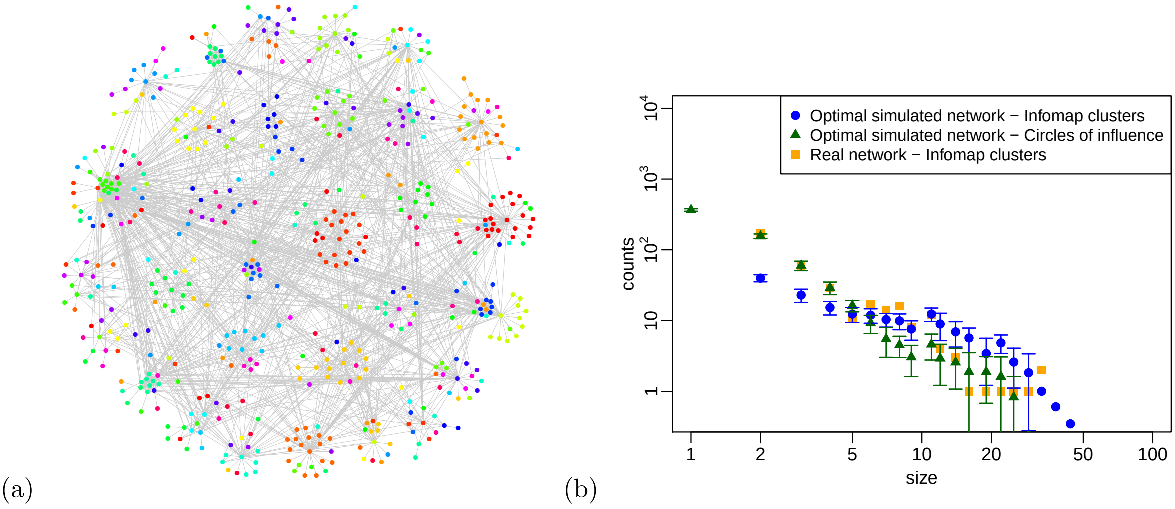


Figure C.10: Electronic components R&D network (SIC code 367). (a) Visual representation of one realization of the optimal simulated network. (b) Size distribution of i. the circles of influence in all realizations of the optimal simulated network, ii. the Infomap clusters in all realizations of the optimal simulated network and iii. the Infomap clusters in the empirical network.

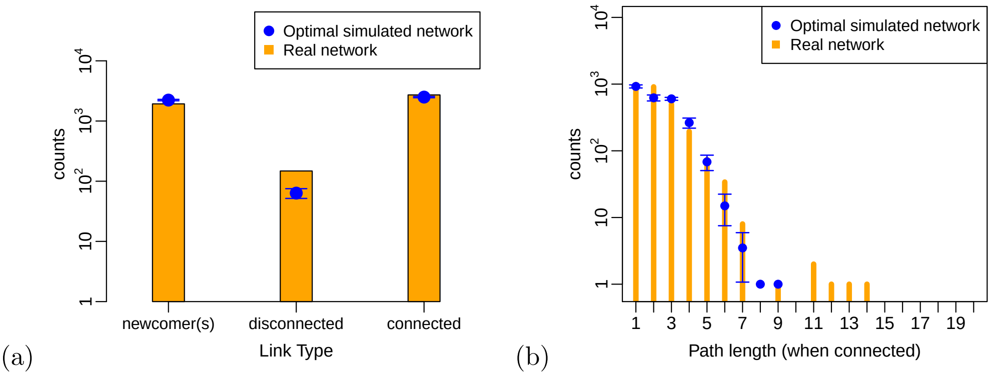


Figure C.11: Electronic components R&D network (SIC code 367). (a) Distribution of link types for empirical and simulated networks. (b) Distribution of path lengths at the moment of link formation (only for nodes belonging to the same connected component).

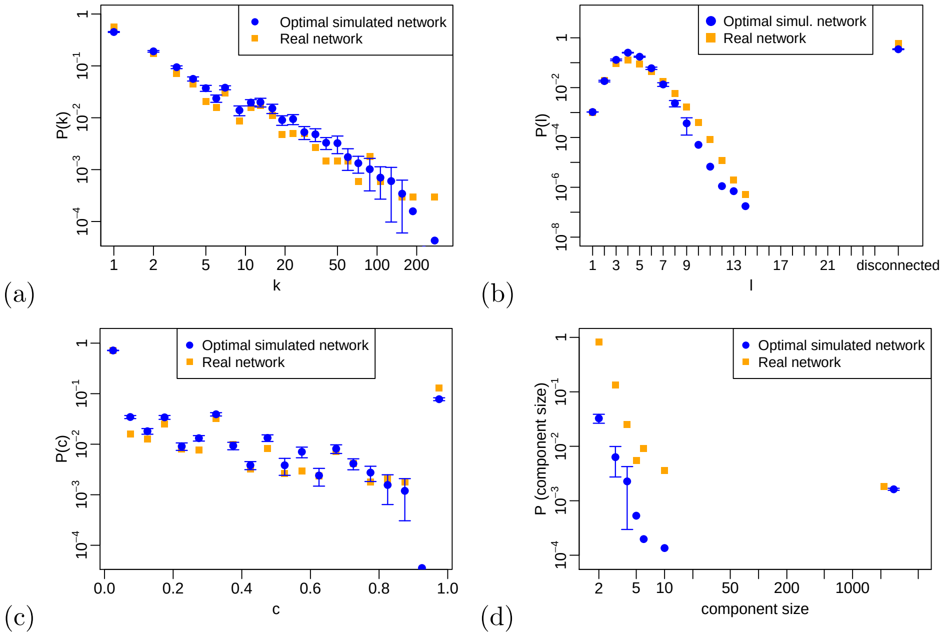


Figure C.12: Computer software R&D network (SIC code 737). Distributions of node degrees (a), path lengths (b), local clustering coefficients (c) and component sizes (d) for the real and the optimal simulated networks.

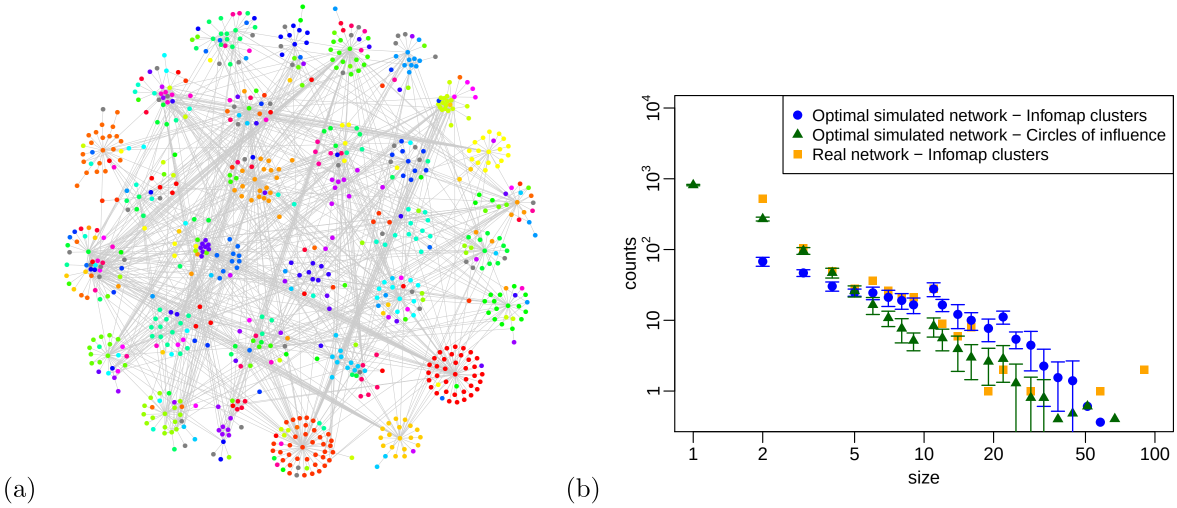


Figure C.13: Computer software R&D network (SIC code 737). (a) Visual representation of one realization of the optimal simulated network. (b) Size distribution of i. the circles of influence in all realizations of the optimal simulated network, ii. the Infomap clusters in all realizations of the optimal simulated network and iii. the Infomap clusters in the empirical network.

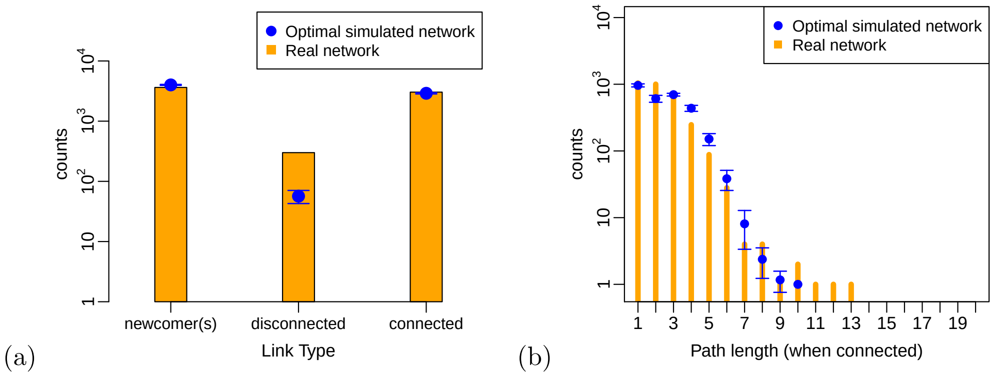
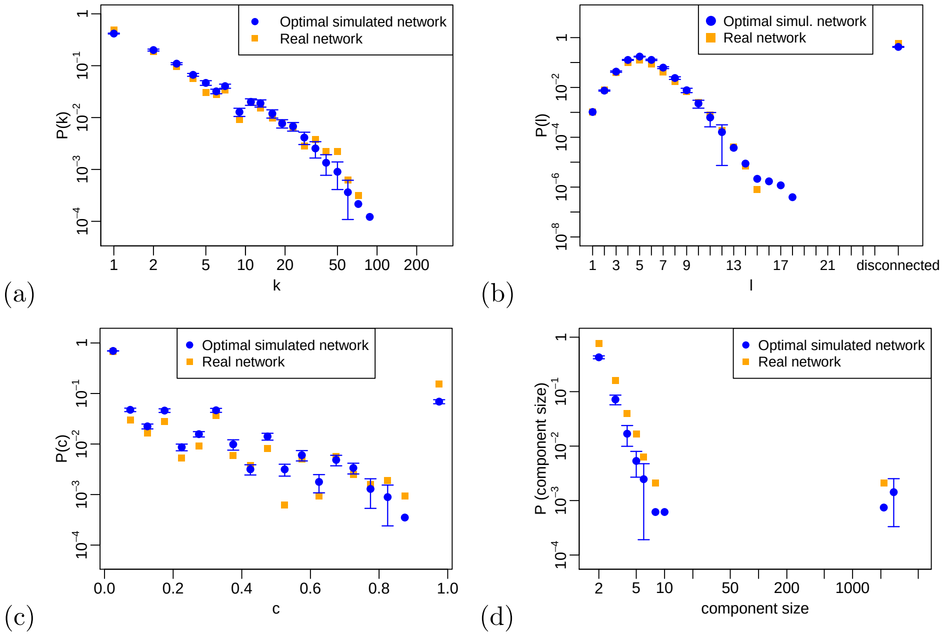
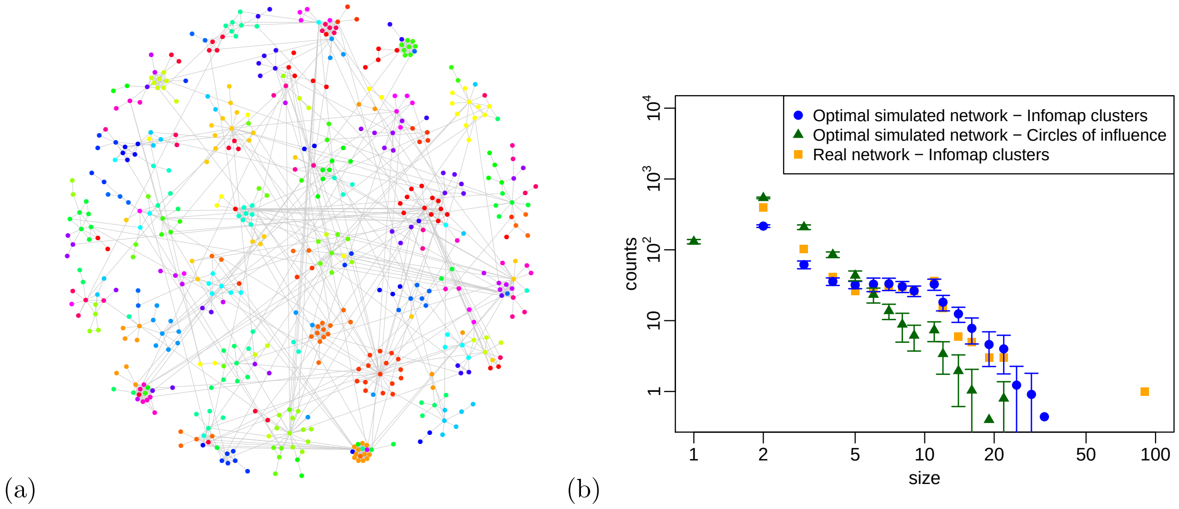


Figure C.14: Computer software R&D network (SIC code 737). (a) Distribution of link types for empirical and simulated networks. (b) Distribution of path lengths at the moment of link formation (only for nodes belonging to the same connected component).



(a) (b) (c) (d)

Figure C.15: R&D, laboratory and testing R&D network (SIC code 873). Distributions of node degrees (a), path lengths (b), local clustering coefficients (c) and component sizes (d) for the real and the optimal simulated networks.



(a) (b)

Figure C.16: R&D, laboratory and testing R&D network (SIC code 873). (a) Visual representation of one realization of the optimal simulated network. (b) Size distribution of i. the circles of influence in all realizations of the optimal simulated network, ii. the Infomap clusters in all realizations of the optimal simulated network and iii. the Infomap clusters in the empirical network.

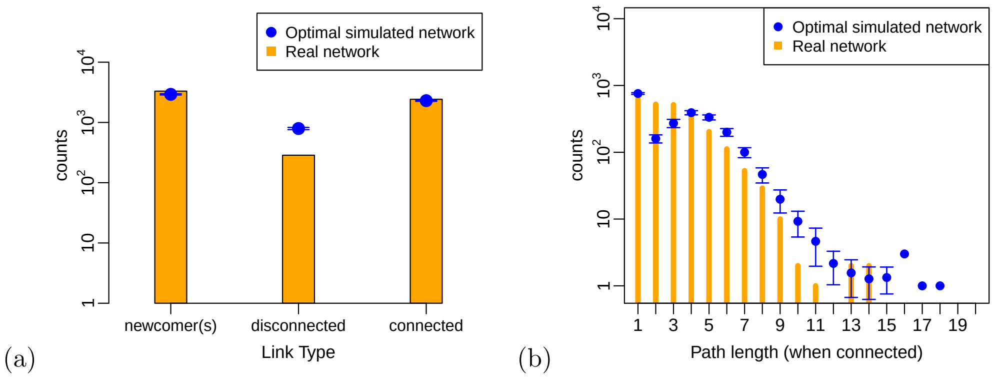


Figure C.17: R&D, laboratory and testing R&D network (SIC code 873). (a) Distribution of link types for empirical and simulated networks. (b) Distribution of path lengths at the moment of link formation (only for nodes belonging to the same connected component).

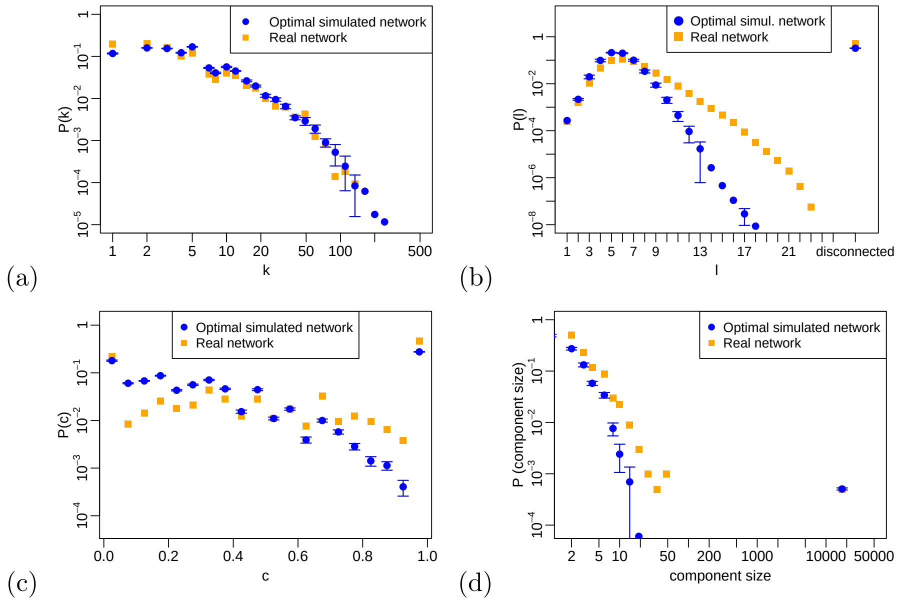


Figure C.18: Quantum mechanics, field theories and special relativity co-authorship network (PACS number 03). Distributions of node degrees (a), path lengths (b), local clustering coefficients (c) and component sizes (d) for the real and the optimal simulated networks.

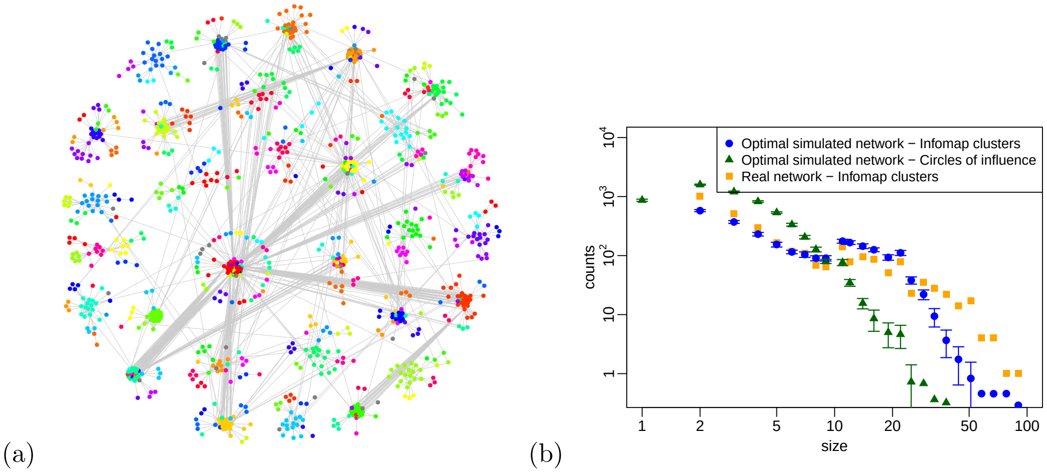


Figure C.19: Quantum mechanics, field theories and special relativity co-authorship network (PACS number 03). (a) Visual representation of one realization of the optimal simulated network. (b) Size distribution of i. the circles of influence in all realizations of the optimal simulated network, ii. the Infomap clusters in all realizations of the optimal simulated network and iii. the Infomap clusters in the empirical network.

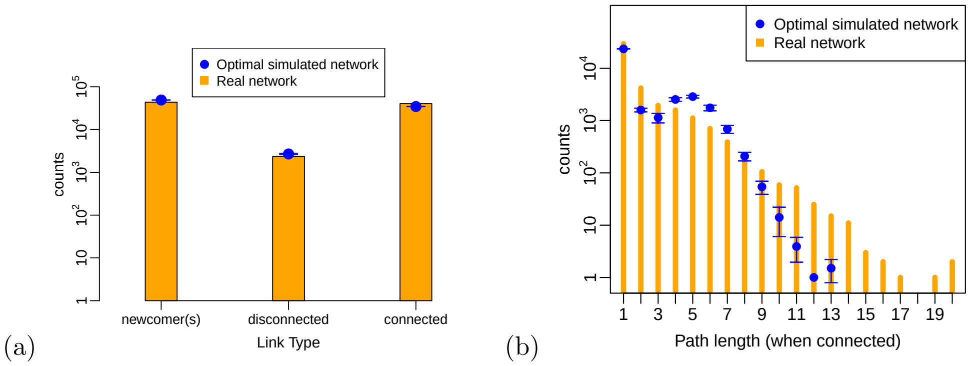
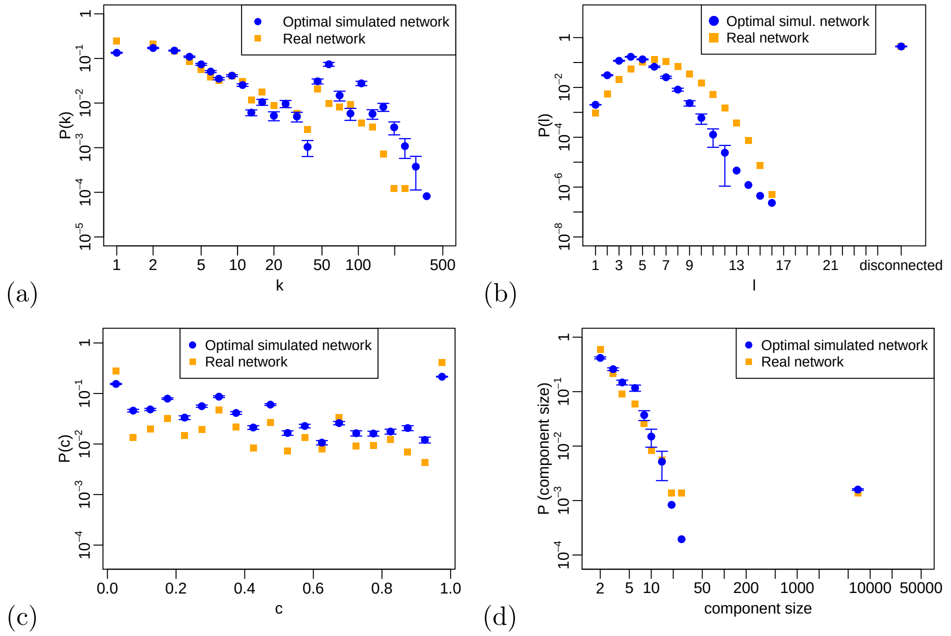


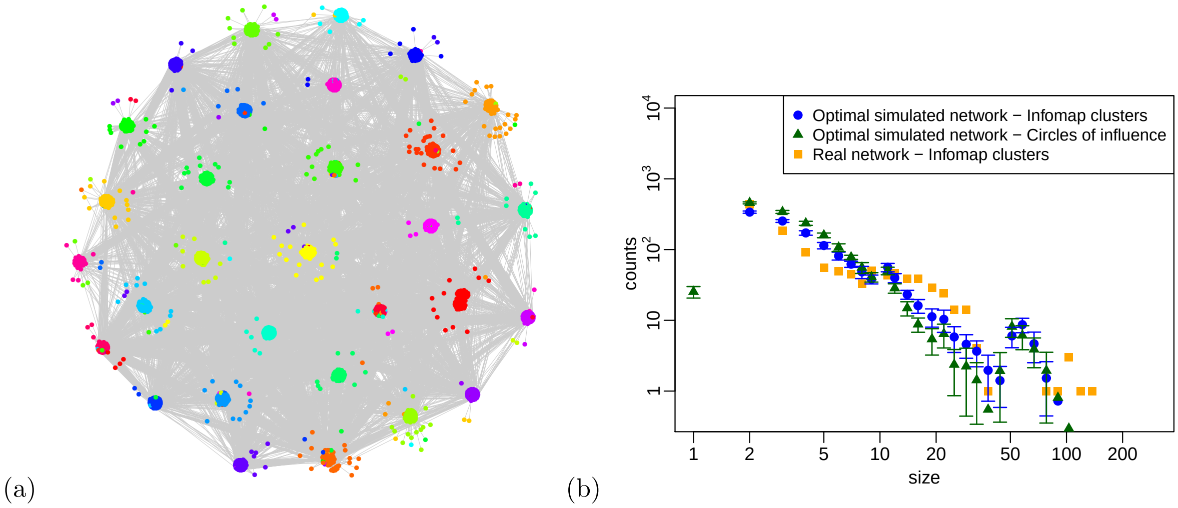
Figure C.20: Quantum mechanics, field theories and special relativity co-authorship network (PACS number 03). (a) Distribution of link types for empirical and simulated networks. (b) Distribution of path lengths at the moment of link formation (only for nodes belonging to the same connected component).



(a) (b)

(c) (d)

Figure C.21: General relativity and gravitation co-authorship network (PACS number 04). Distributions of node degrees (a), path lengths (b), local clustering coefficients (c) and component sizes (d) for the real and the optimal simulated networks.



(a) (b)

Figure C.22: General relativity and gravitation co-authorship network (PACS number 04). (a) Visual representation of one realization of the optimal simulated network. (b) Size distribution of i. the circles of influence in all realizations of the optimal simulated network, ii. the Infomap clusters in all realizations of the optimal simulated network and iii. the Infomap clusters in the empirical network.

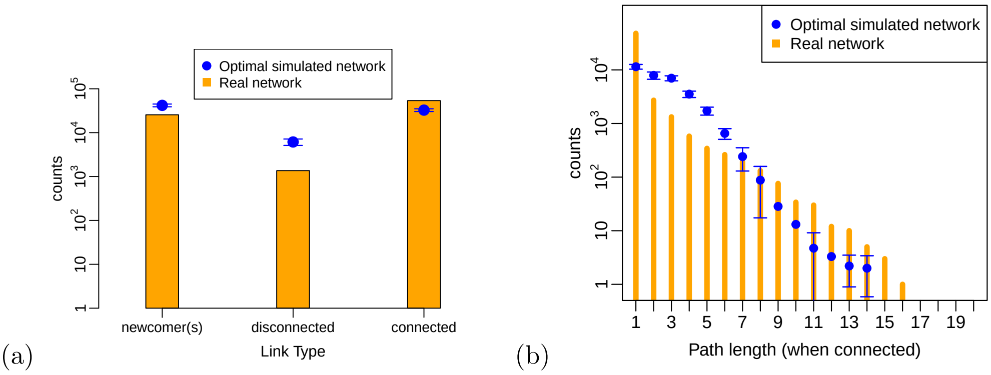


Figure C.23: General relativity and gravitation co-authorship network (PACS number 04). (a) Distribution of link types for empirical and simulated networks. (b) Distribution of path lengths at the moment of link formation (only for nodes belonging to the same connected component).

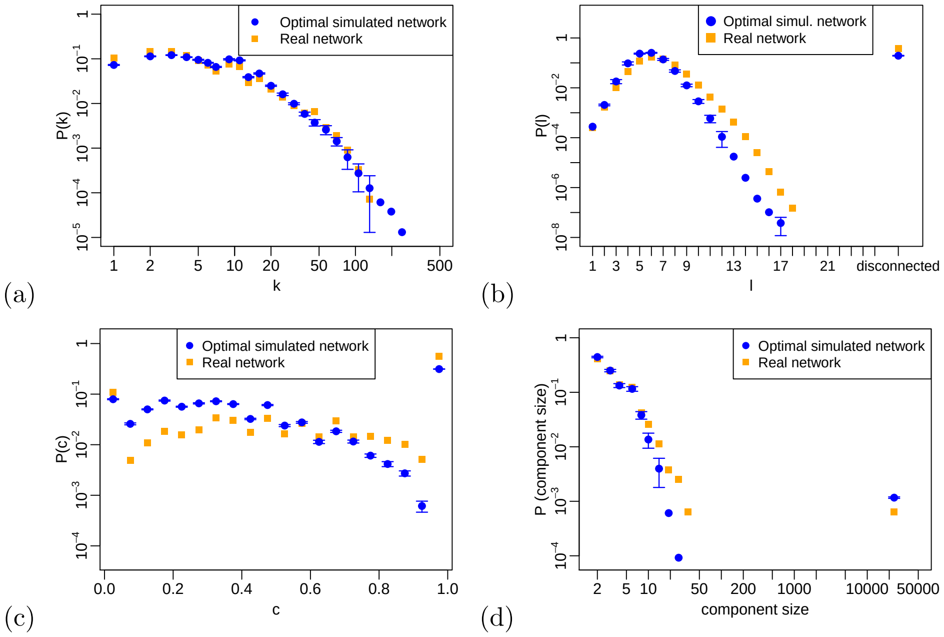


Figure C.24: Optics co-authorship network (PACS number 42). Distributions of node degrees (a), path lengths (b), local clustering coefficients (c) and component sizes (d) for the real and the optimal simulated networks.

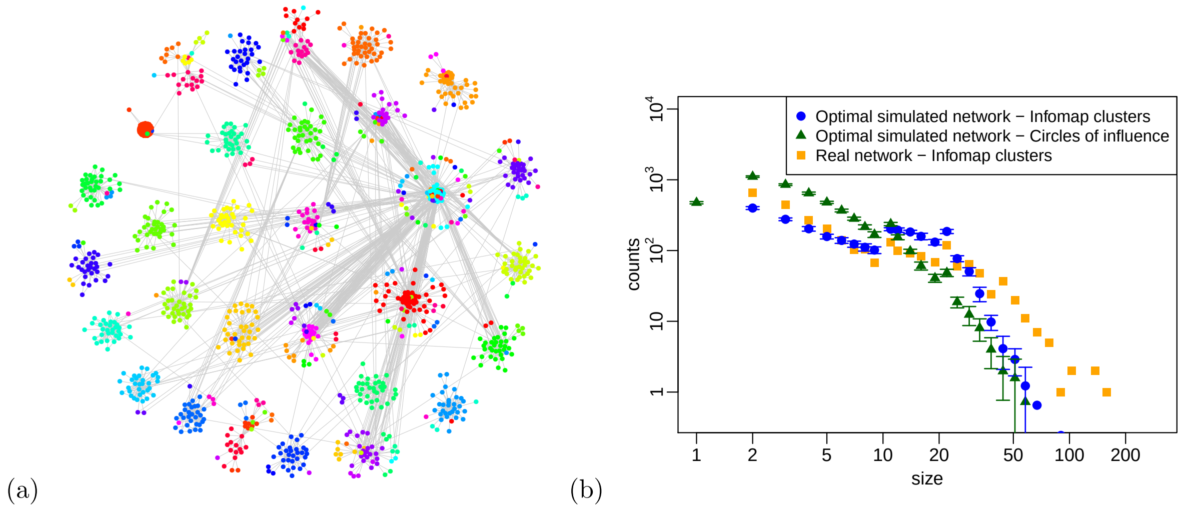


Figure C.25: Optics co-authorship network (PACS number 42). (a) Visual representation of one realization of the optimal simulated network. (b) Size distribution of i. the circles of influence in all realizations of the optimal simulated network, ii. the Infomap clusters in all realizations of the optimal simulated network and iii. the Infomap clusters in the empirical network.

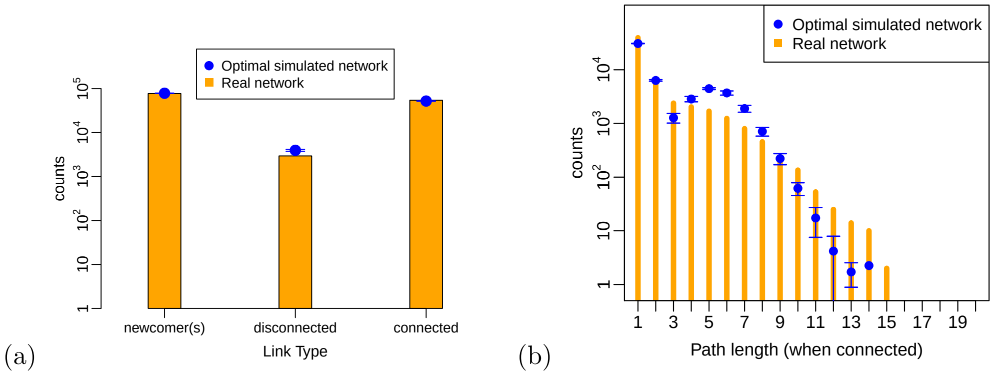


Figure C.26: Optics co-authorship network (PACS number 42). (a) Distribution of link types for empirical and simulated networks. (b) Distribution of path lengths at the moment of link formation (only for nodes belonging to the same connected component).

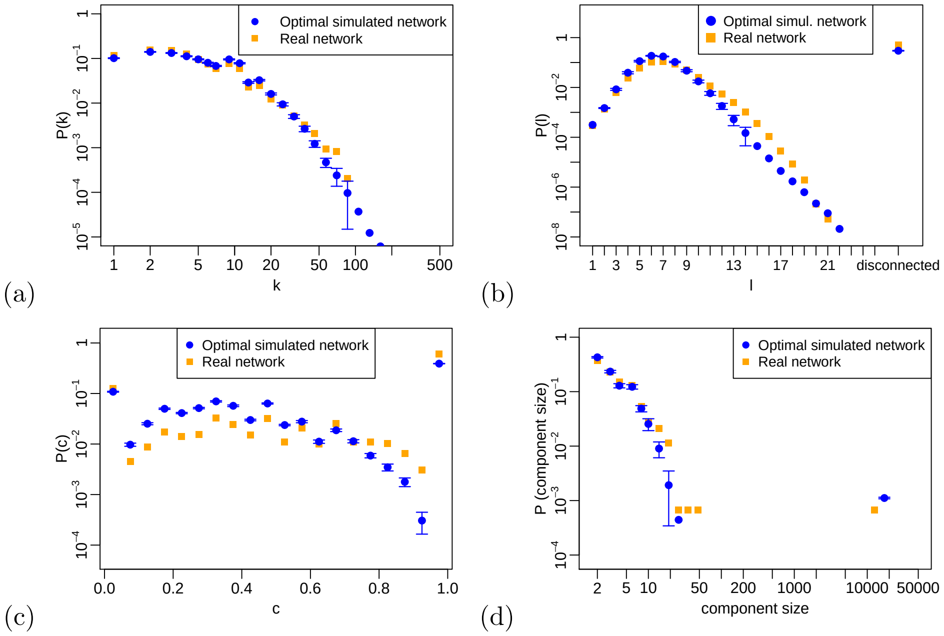


Figure C.27: Electronic transport in condensed matter co-authorship network (PACS number 72). Distributions of node degrees (a), path lengths (b), local clustering coefficients (c) and component sizes (d) for the real and the optimal simulated networks.

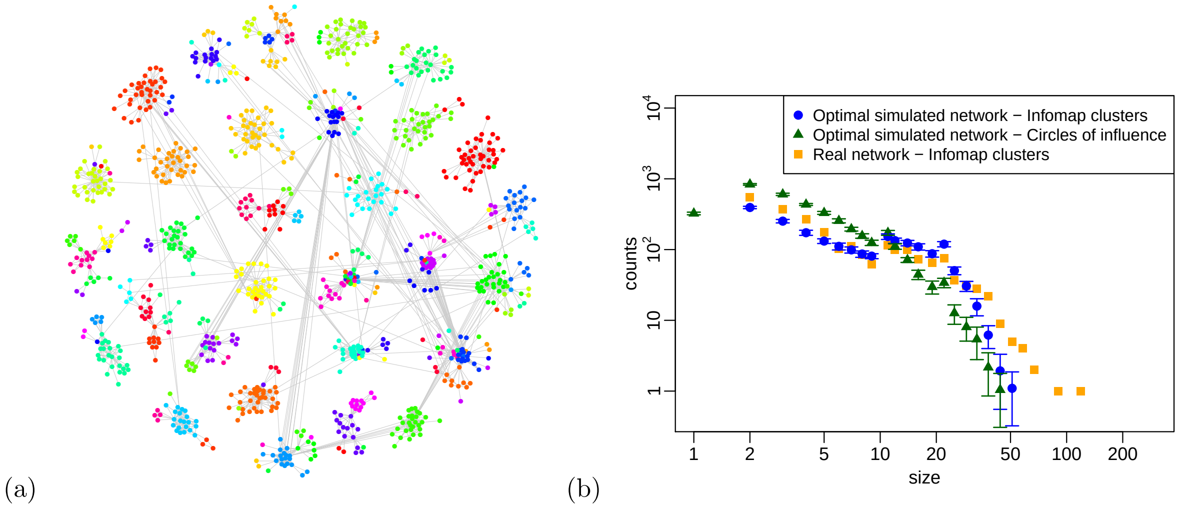
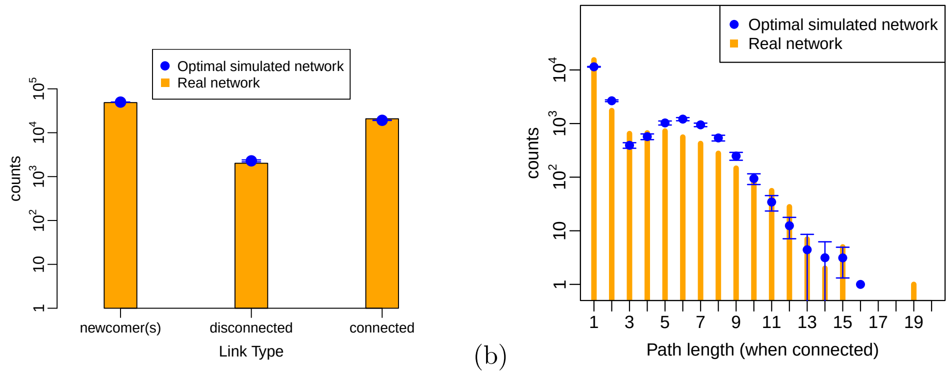
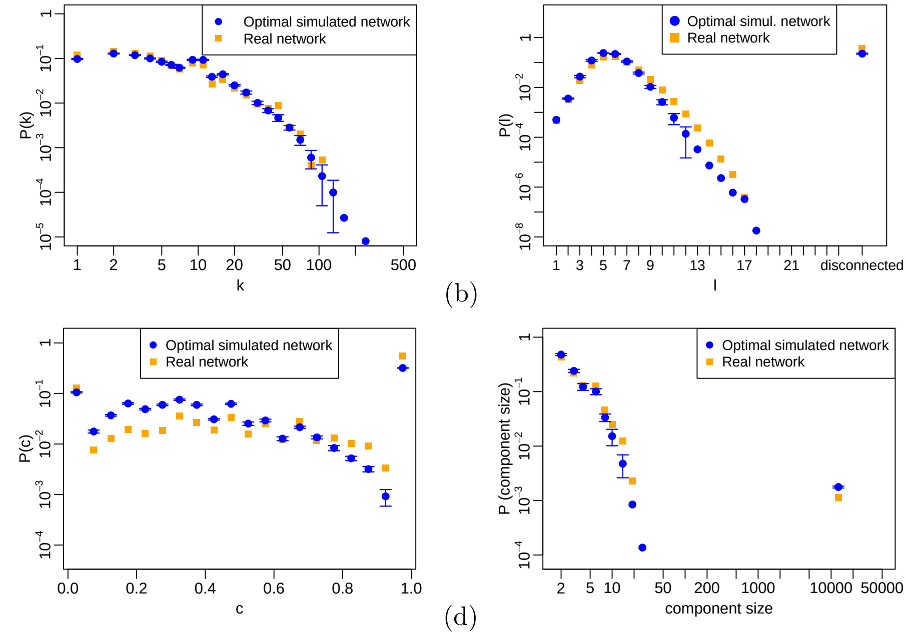


Figure C.28: Electronic transport in condensed matter co-authorship network (PACS number 72). (a) Visual representation of one realization of the optimal simulated network. (b) Size distribution of i. the circles of influence in all realizations of the optimal simulated network, ii. the Infomap clusters in all realizations of the optimal simulated network and iii. the Infomap clusters in the empirical network.



(a) (b)

Figure C.29: Electronic transport in condensed matter co-authorship network (PACS number 72). (a) Distribution of link types for empirical and simulated networks. (b) Distribution of path lengths at the moment of link formation (only for nodes belonging to the same connected component).



(a) (b)

(c) (d)

Figure C.30: Superconductivity co-authorship network (PACS number 74). Distributions of node degrees (a), path lengths (b), local clustering coefficients (c) and component sizes (d) for the real and the optimal simulated networks.

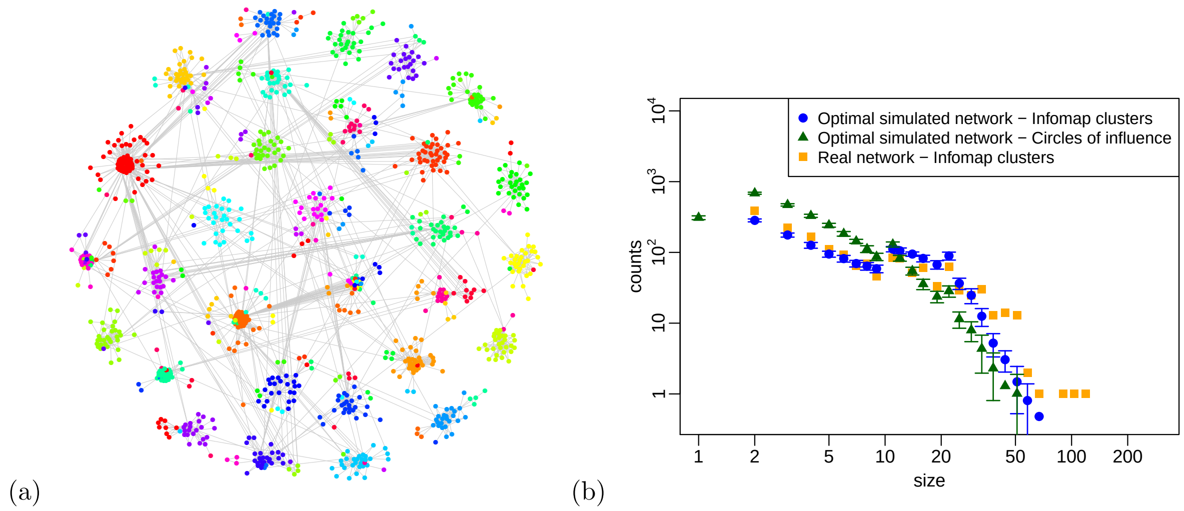


Figure C.31: Superconductivity co-authorship network (PACS number 74). (a) Visual representation of one realization of the optimal simulated network. (b) Size distribution of i. the circles of influence in all realizations of the optimal simulated network, ii. the Infomap clusters in all realizations of the optimal simulated network and iii. the Infomap clusters in the empirical network.

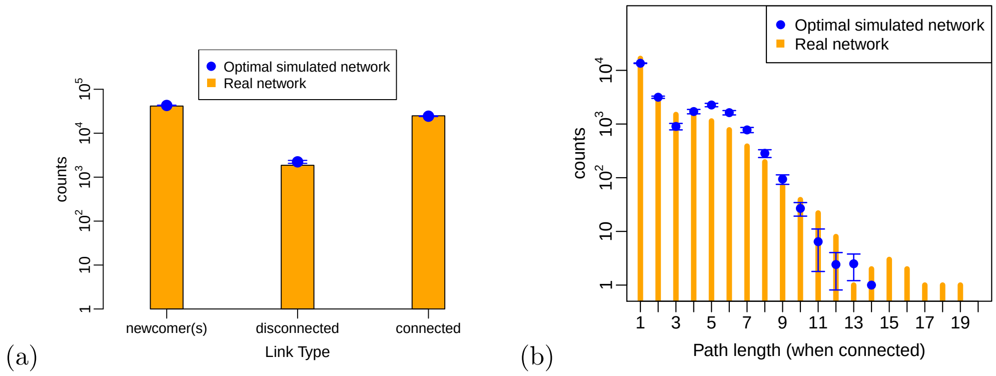


Figure C.32: Superconductivity co-authorship network (PACS number 74). (a) Distribution of link types for empirical and simulated networks. (b) Distribution of path lengths at the moment of link formation (only for nodes belonging to the same connected component).

Appendix D

Supplementary material to Chapter 5

Mean knowledge path We report in Fig. D.1 the mean knowledge path of the collaboration network $\langle K \rangle$ as a function of the dynamics parameter λ and the static parameter ε, for a representative network of $N = 200$ agents moving in a knowledge space with $D = 10$ dimensions.

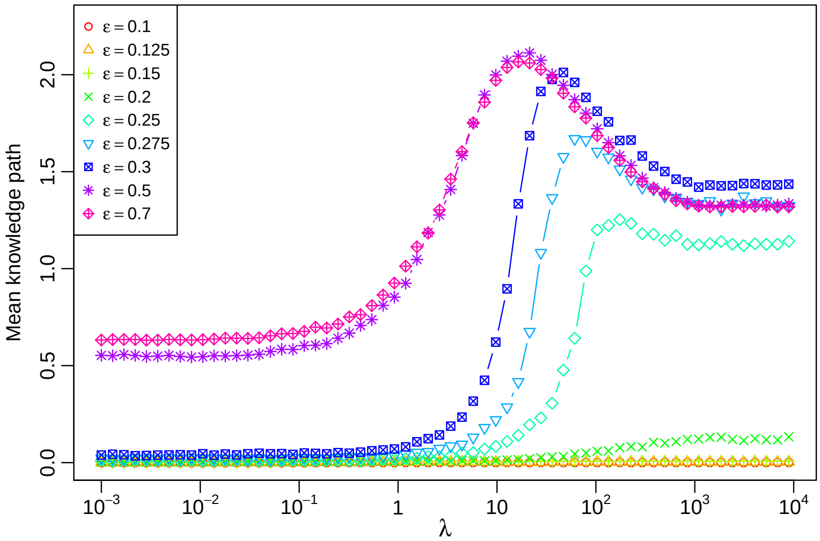


Figure D.1: Mean knowledge path $\langle K \rangle$, as a function of the rewiring rate λ and (parametrically) the interaction radius ε. The R&D network under examination has $N = 200$ nodes and learning rate $\mu = 1$, in a $10-$dimensional knowledge space. We generate 1000 simulations for each parameter set and then average the results.

Appendix E

Supplementary material to Chapter 6

Empirical features of the pooled R&D network with patents We report in Fig. E.1 and Fig. E.2 the activity distribution and the alliance size distribution, respectively, for the pooled R&D network with patent data.

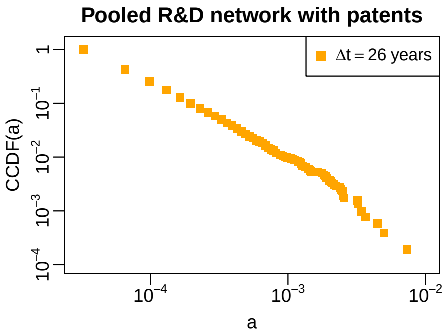


Figure E.1: Complementary cumulative distribution function (CCDF) of the empirical firm activities in the pooled R&D network with patent data, measured on the SDC dataset with 4 different time windows Δt of 1, 5, 10 and 26 years. When the time window is shorter than 26 years (the entire dataset observation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

Numerical simulation results For each of the 324,900 computer simulations we run, we test the resulting generated R&D network with respect to three properties: average degree $\langle k \rangle$, average path length $\langle l \rangle$ and global clustering coefficient C. In Fig. E.3 we show how these three quantities are distributed across all the 324,900 realizations and we compare them with the observed values $\langle k \rangle^{OBS}$, $\langle l \rangle^{OBS}$ and C^{OBS}.

Increasing the dimensionality of the knowledge space For our general modeling framework, we have tested a patent categorization at the “class” level (i.e. the first letter plus two digits of the IPC code). This way, we have obtained a total of 74 classes in

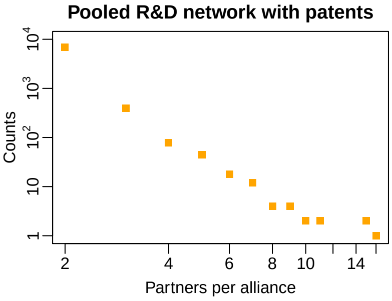


Figure E.2: Distribution of the number of partners per alliance, as measured from the SDC alliance dataset, for the pooled R&D network with patent data.

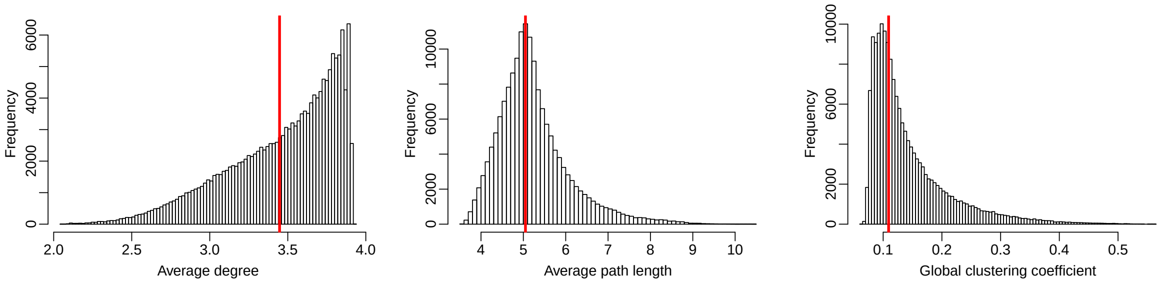


Figure E.3: Distributions of average degree $\langle k \rangle$, average path length $\langle l \rangle$ and global clustering coefficient C across all 324,900 runs of our model (each of the 3,249 points in the parameter space has been explored 100 times). The vertical red lines represent the observed values $\langle k \rangle^{OBS}$, $\langle l \rangle^{OBS}$ and C^{OBS} in the empirical R&D network with patent data.

our metric knowledge space. We find that the computational burden of operating in a 74-dimensional space does not lead to any significant change in our results, if compared to the 7-dimensional knowledge space that we have studied in Chapter 6.

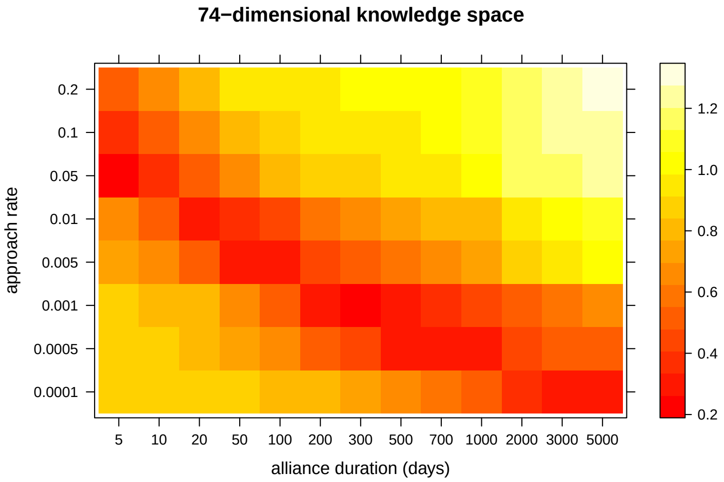


Figure E.4: Goodness score for every point in the parameter space, depicted by means of a heatmap. The color scale corresponds to the score value; the lower the score, the closer the simulated R&D network is to the empirical one. The dimensionality of the knowledge space is $D = 74$, obtained by using a 3-digit IPC patent classification.

List of Figures

List of Tables

Bibliography

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

Ahuja, G. (2000b). The duality of collaboration: Inducements and opportunities in the formation of interfirm linkages. *Strategic management journal* **21(3)**, 317–343.

Albert, R.; Albert, I.; Nakarado, G. L. (2004). Structural vulnerability of the North American power grid. *Physical review E* **69(2)**, 025103.

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

Banks, D. L.; Carley, K. M. (1996). Models for network evolution. *The Journal of Mathematical Sociology* **21(1-2)**, 173–196.

Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature* **435(7039)**, 207–211.

Barabasi, A.-L.; Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509–512.

Bascompte, J.; Jordano, P.; Melián, C. J.; Olesen, J. M. (2003). The nested assembly of plant–animal mutualistic networks. *Proceedings of the National Academy of Sciences* **100(16)**, 9383–9387.

Battiston, S.; Caldarelli, G.; Georg, C. P.; May, R.; Stiglitz, J. (2013). Complex derivatives. *Nature Physics* **9(3)**, 123– 125.

Battiston, S.; Delli Gatti, D.; Gallegati, M.; Greenwald, B.; Stiglitz, J. E. (2012). Liaisons dangereuses: Increasing connectivity, risk sharing, and systemic risk. *Journal of Economic Dynamics and Control* **36(8)**, 1121–1141.

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

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

Bollen, J.; Van de Sompel, H.; Hagberg, A.; Bettencourt, L.; Chute, R.; Rodriguez, M. a.; Balakireva, L. (2009). Clickstream data yields high-resolution maps of science. *PloS one* **4(3)**, e4803.

Borgatti, S. (2005). Centrality and network flow. *Social Networks* **27(1)**, 55–71.

Börner, K.; Chen, C.; Boyack, K. W. (2003). Visualizing knowledge domains. *Annual Review of Information Science and Technology* **37(1)**, 179–255.

Börner, K.; Maru, J. T.; Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 5266–5273.

Boyack, K. W.; Klavans, R.; Börner, K. (2005). Mapping the backbone of science. *Scientometrics* **64(3)**, 351–374.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* **52(3)**, 345–370.

Burkholz, R. (2014). Systemic risk as emerging phenomenon. *Unpublished material.*

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

Caldarelli, G.; Chessa, A.; Pammolli, F.; Gabrielli, A.; Puliga, M. (2013). Reconstructing a credit network. *Nature Physics* **9(3)**, 125– 126.

Callon, M.; Courtial, J.-P.; Turner, W. a.; Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information* **22(2)**, 191–235.

Cantner, U.; Graf, H. (2006). The network of innovators in Jena: an application of social network analysis. *Research Policy* **35(4)**, 463–480.

Clauset, A.; Shalizi, C. R.; Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review* **51(4)**, 661–703.

Cohen, W. (1995). *Handbook of the Economics of Innovation and Technological Change*, Blackwell Publishers Ltd., chap. Empirical Studies of Technological Activity.

Cohen, W.; Levinthal, D. (1990). Absorptive capacity: a new perspective on learning and innovation. *Administrative science quarterly* **35**, 128–152.

Cohen, W. M.; Levinthal, D. A. (1989). Innovation and Learning: The Two Faces of R & D. *The Economic Journal* **99(397)**, 569–596.

Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *The American Journal of Sociology* **94**, 95–120.

Cowan, R.; Jonard, N. (2004). Network structure and the diffusion of knowledge. *Journal of economic Dynamics and Control* **28(8)**, 1557–1575.

Cowan, R.; Jonard, N. (2009). Knowledge portfolios and the organization of innovation networks. *Academy of Management Review* **34(2)**, 320–342.

Cowan, R.; Jonard, N.; Zimmermann, J. (2007). Bilateral collaboration and the emergence of networks. *Management Science* **53(7)**, 1051–1067.

Cox, W. M.; Alm, R. (2008). Creative destruction. *The Concise Encyclopedia of Economics* , Available online at `http://www.econlib.org/library/Enc/CreativeDestruction.html`.

Csardi, G.; Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems* **1695(5)**.

Danon, L.; Diaz-Guilera, A.; Duch, J.; Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment* **2005(09)**, P09008.

Deeds, D.; Hill, C. (1999). An examination of opportunistic action within research alliances-The analysis of discrete structural alternatives. *Journal of Business Venturing* **14(2)**, 141–163.

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

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

Ebers, M. (1997). *The formation of inter-organizational networks*, Oxford University Press, chap. Explaining Inter-Organizational Network Formation. pp. 3–40.

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

Fleming, L.; King, C.; Juda, A. I. (2007). Small Worlds and Regional Innovation. *Organization Science* **18(6)**, 938–954.

Fleming, L.; Marx, M. (2006). Managing Creativity in Small Worlds. *California Management Review* **48(4)**, 6–27.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports* **486**, 75–174.

Fruchterman, T.; Reingold, E. (1991). Graph Drawing by Force-directed Placement. *Software- Practice and Experience* **21(11)**, 1129–1164.

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

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

Garcia, D.; Abisheva, A.; Schweighofer, S.; Serdult, U.; Schweitzer, F. (2015). Network polarization in online politics participatory media. *To appear in Policy and Internet* , 1– 34.

Garcia, D.; Mavrodiev, P.; Schweitzer, F. (2013). Social resilience in online communities: The autopsy of friendster. In: *Proceedings of the first ACM conference on Online social networks.* ACM, pp. 39–50.

Garcia, D.; Tessone, C. J.; Mavrodiev, P.; Perony, N. (2014a). The digital traces of bubbles: Feedback cycles between socio-economic signals in the Bitcoin economy. *Journal of the Royal Society Interface* **11(99)**, 20140623.

Garcia, D.; Weber, I.; Garimella, R. V. K. (2014b). Gender Asymmetries in Reality and Fiction : The Bechdel Test of Social Media. In: *International AAAI Conference on Weblogs and Social Media.* pp. 131–140.

Gersbach, H.; Schmutzler, A. (2003). Endogenous spillovers and incentives to innovate. *Economic Theory* **21(1)**, 59–79.

Gersbach, H.; Schneider, M. T.; Schneller, O. (2013). Basic research, openness, and convergence. *Journal of Economic Growth* **18(1)**, 33–68.

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

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

Gonzalez, M. C.; Hidalgo, C. A.; Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature* **453(7196)**, 779–782.

Goyal, S.; Joshi, S. (2003). Networks of collaboration in oligopoly. *Games and Economic behavior* **43(1)**, 57–85.

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

Granovetter, M. (1985). Economic Action and Social Structure: The Problem of Embeddedness. *The American Journal of Sociology* **91(3)**, 481–510.

Granovetter, M. S. (1973). The Strength of Weak Ties. *The American Journal of Sociology* **78**, 1360–1380.

Granovetter, M. S. (1983). The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory* **1**, 201–233.

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

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

Guimera, R.; Uzzi, B.; Spiro, J.; Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science* **308(5722)**, 697–702.

Gulati, R. (1995a). Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances. *Academy of management journal* **38(1)**, 85–112.

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

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

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

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

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

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

Hill, B. M. (1975). A Simple General Approach to Inference About the Tail of a Distribution. *The Annals of Statistics* **3(5)**, pp. 1163–1174.

Holme, P. (2005). Core-periphery organization of complex networks. *Phys. Rev. E* **72**, 046111.

Holme, P.; Saramäki, J. (2012). Temporal networks. *Physics reports* **519(3)**, 97–125.

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

Jackson, M. O.; Rogers, B. W. (2007). Meeting Strangers and Friends of Friends: How Random Are Social Networks? *American Economic Review* **97(3)**, 890–915.

Jackson, M. O.; Wolinsky, A. (1996). A Strategic Model of Social and Economic Networks. *Journal of Economic Theory* **71(1)**, 44–74.

Jaffe, A.; Trajtenberg, M. (2002). *Patents, Citations, and Innovations: A Window on the Knowledge Economy.* MIT Press.

Kahneman, D.; Tversky, A. (1996). On the reality of cognitive illusions. *Psychological review* **103(3)**, 582–91.

Kaushik, R.; Battiston, S. (2013). Credit default swaps drawup networks: Too interconnected to be stable? *PloS one* **8(7)**, e61815.

Kleinberg, J.; Suri, S.; Tardos, É.; Wexler, T. (2008). Strategic network formation with structural holes. In: *Proceedings of the 9th ACM Conference on Electronic Commerce.* ACM, pp. 284–293.

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

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

König, M. D.; Tessone, C. J.; Zenou, Y. (2010). From assortative to dissortative networks: the role of capacity constraints. *Advances in Complex Systems* **13(04)**, 483–499.

König, M. D.; Tessone, C. J.; Zenou, Y. (2014). Nestedness in Networks: A Theoretical Model and Some Applications. *Theoretical Economics* **8**, Accepted, to appear, forthcoming in Theoretical Economics.

Lancichinetti, A.; Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Phys. Rev. E* **80**, 056117.

Lazer, D.; Friedman, A. (2007). The Network Structure of Exploration and Exploitation. *Administrative Science Quarterly* **52(4)**, 667–694.

Letterie, W.; Hagedoorn, J.; van Kranenburg, H.; Palm, F. (2008). Information gathering through alliances. *Journal of Economic Behavior & Organization* **66(2)**, 176–194.

Leydesdorff, L. (1987). Various Methods for the Mapping of Science. *Scientometrics* **11**, 291–320.

Lissoni, F.; Llerena, P.; Sanditov, B. (2013). Small Worlds in Networks of Inventors and the Role of Academics: An Analysis of France. *Industry and Innovation* **20(3)**, 195–220.

Liu, X.; Bollen, J.; Nelson, M. L.; Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Information processing & management* **41(6)**, 1462–1480.

McCain, K. W. (1991). Mapping Economics through the Journal Literature : An Experiment in Journal Cocitation Analysis. *Journal of the American Society for Information Science* **42(4)**, 290–296.

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

Nelson, R. R.; Winter, S. G. (1982). *An evolutionary theory of economic change.* Cambridge, Mass. : Belknap Press of Harvard University Press.

Newman, M. (2004a). Detecting community structure in networks. *The European Physical Journal B-Condensed Matter and Complex Systems* **38(2)**, 321–330.

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

Newman, M. E. J. (2002). Assortative Mixing in Networks. *Physical Review Letters* **89(20)**, 208701.

Newman, M. E. J. (2003). Mixing patterns in networks. *Physical Review E* **67(2)**, 026126.

Newman, M. E. J. (2004b). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 5200–5205.

Newman, M. E. J.; Girvan, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* **69**, 026113.

Nonaka, I.; Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation.* Oxford university press.

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

Pastor-Satorras, R.; Vazquez, A.; Vespignani, A. (2001). Dynamical and Correlation Properties of the Internet. *Physical Review Letters* **87**.

Pastor-Satorras, R.; Vespignani, A. (2007). *Evolution and structure of the Internet: A statistical physics approach.* Cambridge University Press.

Perra, N.; Goncalves, B.; Pastor-Satorras, R.; Vespignani, A. (2012). Activity driven modeling of time varying networks. *Scientific Reports* **2**, 469.

Pfitzner, R.; Scholtes, I.; Garas, A.; Tessone, C. J.; Schweitzer, F. (2013). Betweenness preference: Quantifying correlations in the topological dynamics of temporal networks. *Physical review letters* **110(19)**, 198701.

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

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

Powell, W.; Grodal, S. (2006). *Oxford Handbook of Innovation*, Oxford University Press, USA, chap. Networks of Innovators.

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

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

Price, D. J. D. (1965). Networks of scientific papers. *Science* **149**, 510–515.

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

Ramasco, J. J.; Dorogovtsev, S. N.; Pastor-Satorras, R. (2004). Self-organization of collaboration networks. *Physical Review E* **70(3)**, 036106.

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

Reed, W. (2001). The Pareto, Zipf and other power laws. *Economics Letters* **74(1)**, 15–19.

Reichardt, J.; Bornholdt, S. (2006). When are networks truly modular? *Physica D: Nonlinear Phenomena* **224(1)**, 20–26.

Ribeiro, B. (2014). Modeling and Predicting the Growth and Death of Membership-based Websites. In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14, New York, NY, USA: ACM, pp. 653–664.

Riccaboni, M.; Pammolli, F. (2002). On firm growth in networks. *Research Policy* **31(8-9)**, 1405–1416.

Rodriguez-Girones, M. A.; Santamaria, L. (2006). A new algorithm to calculate the nestedness temperature of presence-absence matrices. *Journal of Biogeography* **33(5)**, 924–935.

Roijakkers, N.; Hagedoorn, J. (2006). Inter-firm R&D partnering in pharmaceutical biotechnology since 1975: Trends, patterns, and networks. *Research Policy* **35(3)**, 431–446.

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

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

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

Rosvall, M.; Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* **105(4)**, 1118–1123.

Rowley, T.; Behrens, D.; Krachhardt, D. (2000). Redundant Governance Structures: An Analysis of Structural and Relational Embeddness in the Steel and Semiconductor Industries. *Strategic Management Journal* **21(3)**, 369–386.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika* **31(4)**, 581–603.

Salter, A. J.; Martin, B. R. (2001). The economic benefits of publicly funded basic research: a critical review. *Research policy* **30(3)**, 509–532.

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

Sarigöl, E.; Pfitzner, R.; Scholtes, I.; Garas, A.; Schweitzer, F. (2014). Predicting scientific success based on coauthorship networks. *EPJ Data Science* **3(1)**, 9.

Schilling, M. (2009). Understanding the Alliance Data. *Strategic Management Journal* **30**, 233–260.

Scholtes, I.; Wider, N.; Pfitzner, R.; Garas, A.; Tessone, C. J.; Schweitzer, F. (2014). Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks. *Nature communications* **5**, 5024.

Schumpeter, J. A. (1942). *Capitalism, Socialism and Democracy.* Harper, New York.

Schweitzer, F. (2007). *Brownian Agents and Active Particles.* Springer Series in Synergetics, Springer Berlin Heidelberg.

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

Schweitzer, F.; Fagiolo, G.; Sornette, D.; Vega-Redondo, F.; Vespignani, A.; White, D. R. (2009). Economic networks: The new challenges. *Science* **325(5939)**, 422–425.

Simon, H. A. (1955). On a Class of Skew Distribution Functions. *Biometrika* **42(3/4)**, 425–440.

Snijders, T. (2001). The statistical evaluation of social network dynamics. *Sociological Methodology* **31(1)**, 361–395.

Suzumura, K. (1992). Cooperative and Noncooperative R&D in an Oligopoly with Spillovers. *The American Economic Review* **82(5)**, 1307–1320.

Tessone, C. J.; Zanette, D. H. (2012). Synchronised firing induced by network dynamics in excitable systems. *EPL (Europhysics Letters)* **99(6)**, 68006.

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

Tiwana, A. (2008). Do bridging ties complement strong ties? An empirical examination of alliance ambidexterity. *Strategic Management Journal* **29(3)**, 251–272.

Uzzi, B.; Amaral, L. A.; Reed-Tsochas, F. (2007). Small-world networks and management science research: a review. *European Management Review* **4(2)**, 77–91.

Vazquez, F.; Zanette, D. (2010). Epidemics and chaotic synchronization in recombining monogamous populations. *Physica D: Nonlinear Phenomena* **239(19)**, 1922–1928.

Vega-Redondo, F.; Goyal, S. (2007). Structural holes in social networks. *Journal of Economic Theory* **137(1)**, 460–492.

Vespignani Alessandro (2010). Complex networks: The fragility of interdependency. *Nature* **464(7291)**, 984–985.

Veugelers, R. (October 1998). Collaboration in R&D: An Assessment of Theoretical and Empirical Findings. *De Economist* **146**, 419–443(25).

Walker, G. (2005). *Handbook of Strategic Alliances*, Thousand Oaks, CA: Sage, chap. Networks of Strategic Alliances.

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

Watts, D. J.; Strogatz, S. H. (1998). Collective Dynamics of Small-World Networks. *Nature* **393**, 440–442.

Westbrock, B. (2010). Natural concentration in industrial research collaboration. *The RAND Journal of Economics* **41(2)**, 351–371.

Winter, S. G. (1984). Schumpeterian competition in alternative technological regimes. *Journal of Economic Behavior and Organization* **5(3)**, 287–320.

Yang, J.; Leskovec, J. (2012). Defining and evaluating network communities based on ground-truth. In: *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*. ACM, p. 3.

Personal information

Surname(s) / First name(s)	**Tomasello, Mario Vincenzo**
Address(es)	Manessestrasse 66, CH-8003 Zurich (Switzerland)
Email(s)	mtomasello@ethz.ch
Nationality(-ies)	Italian
Date of birth	February 12, 1986

Education and training

March 2011 – present	**Ph.D.** student at the *Chair of Systems Design*, **ETH Zurich**, Department of Management, Technology and Economics (Zurich, Switzerland).
February – August 2013	Seven-month visiting period at **Northeastern University** (Boston, USA).
December 2010	**Master in Economics and Management** (Environment and Energy) at *Bocconi University* (Milano, Italy). 110/110 *cum laude* - Average grade: 28.75/30.
December 2010	Diploma at *Collegio di Milano* (Milano, Italy), recognized by MIUR (*Italian Ministry of Education*) as a college of excellence.
July 2009	**Master of Science in Physics** at *University of Catania* (Catania, Italy). 110/110 *cum laude* - Average grade: 30/30.
November 2008	Diploma at the *Scuola Superiore di Catania* (Catania, Italy), recognized by MIUR (*Italian Ministry of Education*) as a college of excellence. 70/70 *cum laude* - Average grade: 30/30.
July 2007	**Bachelor of Science in Physics** at *University of Catania* (Catania, Italy). 110/110 *cum laude* - Average grade: 29.8/30.

Publications

2014	*The Role of Endogenous and Exogenous Mechanisms in the Formation of R&D Networks* (M.V.Tomasello, N.Perra, C.J.Tessone, M.Karsai, F.Schweitzer), *Scientific Reports*, No.4, 5679 (2014)
2014	*Innovator Networks* (M.V.Tomasello, M.Müller, F.Schweitzer), In: *Encyclopedia of Social Network Analysis and Mining*, Springer, New York, pp. 737–742 (2014)
2014	*Selection rules in alliance formation: Strategic decisions or abundance of choice?* (A.Garas, M.V.Tomasello, F.Schweitzer), *arXiv:1403.3298* (2014)
2013	*The Rise and Fall of R&D Networks* (M.V.Tomasello, M.Napoletano, A.Garas, F.Schweitzer), *arXiv:1304.3623* (2013)
2011	*Analyses of the As doping of SiO$_2$/Si/SiO$_2$ nanostructures* (F.Ruffino, M.V.Tomasello, M.Miritello, R.De Bastiani, G.Nicotra, C.Spinella, M.G.Grimaldi), *Physica Status Solidi C, 8*, No.3, 863 – 866(2011)
2010	*Nanostructuring in Ge by self-ion implantation* (L.Romano, G.Impellizzeri, M.V.Tomasello, F.Giannazzo, C.Spinella, M.G.Grimaldi), *Journal of Applied Physics, 107*, 084314(2010)
2010	*Arsenic Doping of Silicon-based low-dimensional systems* (F.Ruffino, M.V.Tomasello, M.Miritello, G.Nicotra, C.Spinella, M.G.Grimaldi), *Applied Physics Letters, 96*, 093116(2010)

Talks

July 2014	*The Rise and Fall of R&D Networks*, International Schumpeter Society Conference, Jena (Germany)
June 2014	*The Rise and Fall of R&D Networks*, DRUID Society Conference, Copenhagen (Denmark)
March 2014	*The Role of Endogenous and Exogenous Mechanisms in the Formation of R&D Networks*, OFCE-SKEMA Business School, Nice (France)

March 2013 — *An activity-driven model for the growth of R&D networks*, Northeastern University, Boston (U.S.A.)

September 2012 — *Network dynamics and the creation of knowledge in R&D networks*, Latsis Symposium, ETH Zurich, Zurich (Switzerland)

September 2012 — *Network dynamics and the creation of knowledge in R&D networks*, ECCS (European Conference on Complex Systems) 2012, Brussels (Belgium)

June 2012 — *The evolution of R&D networks across industries*, SKEMA Paper Development Workshop 2012, Nice (France)

September 2011 — *Cuttlefish for visualization of network dynamics in research alliances*, ASNA (Application of Social Network Analysis) Conference 2011, Zurich (Switzerland)

Company experiences

October 2010 – October 2011 — Winner of the scholarship *“Leaders of the future”*, by **The European House – Ambrosetti**, an Italian professional consulting group. The scholarship, reserved to the 15 most outstanding university students of the year in Italy, was paid by *ENEL*, the largest Italian electric company.

September – December 2010 — **Accenture** – Internship in Management Consulting (Resources area), Milano (Italy). Project about *Smart Cities*: development of business plans to assess and compare the effectiveness of several *green technologies*.

Additional talks

January 2011 — *New ways of communications with customers using new technologies: Web2.0 and Smartphone*, marketing project in collaboration with **L’Oréal**, a multinational firm (Milano, Italy).

April 2010 — *Statistics of women behaviour and education in Italy*, talk presented at *Integrated Management of Environment* workshop, in collaboration with **TetraPak**, a multinational firm (Reggio Emilia, Italy).

Awards — Title of “Alfiere del Lavoro” (Rome, October 2004), reserved to the **25 most outstanding** high school students of the year in Italy.

High school scientific competitions

May 2003 — First place in the regional phase (region Calabria) of the *Italian Physics Olympics*

May 2004 — First place in the regional phase (region Calabria) of the *Italian Physics Olympics*

May 2004 — Participating in the final phase of the*Italian Physics Olympics*; award-winning with *’menzione di merito’* (mention)

Languages skills

Mother tongue(s) — **Italian**

Self-assessment
European level(*)

	Understanding		**Speaking**		**Writing**
	Listening	Reading	Spoken interaction	Spoken production	
English	C1 Proficient user	C1 Proficient user	C1 Proficient user	C1 Proficient user	C1 Proficient user
German	A1 Basic user	A2 Basic user	A1 Basic user	A1 Basic user	A1 Basic user

(*) *Common European Framework of Reference (CEF) level*

Computer skills — Programming languages C, Java, Python, R; Latex and Emacs; operating systems: Windows, Linux (Ubuntu), Mac OS.

Interests — Music: lead guitarist in a band for two years (2004-2005); soundtrack of the short movie “Miki” (Italy, 2008). Passion for electronics and cars, repairing old vehicles and appliances, do-it-yourself. Amateur actor. Volunteer experience: counselor at three summer camps (2003-2004-2005). Sport: cycling and running.