Understanding the social dimension of knowledge through complex network analysis

Elena Mas Tur

ISBN 978-94-629-5400-7

Published by: Uitgeverij BOXpress | | proefschriftmaken.nl

Understanding the social dimension of knowledge through complex network analysis

Het begrijpen van de sociale dimensie van kennis aan de hand van de analyse van complexe netwerken

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op maandag 23 mei 2016 des ochtends te 10.30 uur

door

Elena María Mas Tur

geboren op 11 augustus 1987

te Valencia, Spanje

Promotor: Prof.dr. K. Frenken

Copromotoren: Dr. J.M. Azagra-Caro

Dr. P. Zeppini

This thesis was accomplished with financial support from the Spanish National Research Council JAE-PreDoc fellowship co-financed by the ESF.

Committee: Prof.dr. R.A. Boschma

Prof.dr.ir. V. Buskens

Prof.dr. F. Alkemade

Prof.dr. B. van Looy

Dr. R. De Langhe

Contents

CHAPTER 1

INTRODUCTION

1.1 Knowledge

Ever since the seminal work on creative destruction by Schumpeter (1934), innovation has been widely accepted to play a central role in economic growth. Over and over, innovation introduces changes in the economic structure via more efficient forms of production or new commodities. In consequence, innovation can drastically alter the conditions under which economic activity takes place and transform the product space or the technological field. Hence, innovation has been at the focus of both policy making and academia (Fagerberg and Verspagen, 2009).

Innovation finds its first and foremost origin in novelty, leading to improvements in the production process or new products. Those improvements can result from new problems arising that need solving, or from new solutions being

found for old problems. In both cases, creative thinking holds the key to the creation of new knowledge and new ideas. Throughout this thesis, we will refer to knowledge as an abstract concept that encompasses ideas, information, theories, procedures, etc.

In present times, knowledge and innovation have experienced a process of evolution towards a highly structured system. Before the industrial revolution, and before the advent of Information and Communication Technologies, knowledge relied on a disconnected universe of islands, and innovation was centered on the figure of the isolated scientist or inventor, often represented in the folklore as an awkward genius or a lonely prospector (Singh and Fleming, 2010). Nowadays, knowledge production is far more distributed over individuals and organizations often referred to as innovation systems.

There are three main reasons behind this historical change in knowledge systems and innovation processes. First, increased communication and mobility have led to an innovation model that is largely based on collaborative efforts, among scientists as well as organizations (Graf, 2006). Nowadays it usually requires a joint effort to produce new knowledge (Boschma, 2005). Every agent specializes in a small region of the knowledge field, and the interaction of many such deep and narrow specializations can open new, original research directions and provide hybrid methodologies to unveil new knowledge. Collaboration is not always straightforward, since agents need to overcome cultural barriers such as differences in their incentives and threats innate to the collaboration process like free-riding. Understanding how collaboration arises to form teams, and how the teams tackle and solve problems is key to explain how knowledge is produced. At the same time, a thorough comprehension of how knowledge is produced is a necessary first step towards improving the performance of the knowledge creation system.

Second, the social implications of innovation have become hugely important, due to the role of markets, governments and other stakeholders in articulating what is to be innovated and when (Nemet, 2009; Walsh, 1984). Knowledge production alone is not enough to trigger innovation and economic growth. Innovation requires not only the invention phase, but also the commercialization stage. A second step is essential in which the knowledge is assimilated by the society. Abstract new ideas crystallize into improvements and newness to the production system, and are integrated in industries or launched in markets.

Third, innovation is now more than ever the outcome of recombination events (Arthur, 2009; van den Bergh, 2008), as different technological capabilities are pulled together and new functionalities created from the combination of different modules. In all these three aspects one can see how the innovation process in recent times has started feeding itself. An efficient diffusion of knowledge, moreover, builds positive reinforcement for the creation of further new knowledge (Cowan and Jonard, 2003; Konig et al., 2011). First of all, as increasing amounts of knowledge disseminate through the society, even more recombinations can be formed to create further knowledge. In addition, increasing the amount of knowledge produced allows for more knowledge spreading through the system. This strengthens the feedback between both processes, resulting in a self-reinforcing virtuous cycle of knowledge production and diffusion.

A particular but fundamental aspect of these changes in the organization of knowledge systems and innovation dynamics is that they have assumed more and more a networked connotation, in many respects: R&D collaborations naturally define a network of inventors (Cowan et al., 2007), innovations often diffuse through social networks of word-of-mouth communication (Alkemade and Castaldi, 2005), and recombinant technologies define and shape an ever complex technological network (Frenken et al., 2012). Furthermore, these three aspects of knowledge and innovation complexity are entangled together. The intangi-

ble nature of knowledge can make knowledge diffusion intricate. Collaboration networks not only favor knowledge creation, they also encourage knowledge diffusion by building trust between its members, establishing commonalities such as shared aims or norms, etc. Moreover, since the truth of knowledge is not necessarily self-evident, its acceptance and subsequent diffusion depends also on the social environment (Kuhn, 1970).

The complex social network nature of knowledge systems and processes is the main focus of this thesis. We will show how network science has provided some crucial tools for the examination of knowledge and innovation networks, with a number of results that add to the literature on this subject. Understanding the role of social networks in the production and diffusion of new knowledge has strong implications for a number of policy relevant issues. Not only this is relevant to innovation policy per se, but also for pressing global problems such as sustainable growth, climate change, and social inequality.

1.2 Networks

Network science was initiated in 1736 with the first definition of a graph to solve the Konisberg bridges problem by Euler. Traditionally, graph theory dealt with regular graphs, where every node has the same number of connections, until Erdos and Renyi (1959) introduced probabilistic properties to the structure of the network. Their model, with a fixed set of nodes connected by edges with equal probability, guided research on networks for several decades. With the advent of computerization of data and increase of computer power, network science experienced a dramatic growth. These included the first advance in network science that relied entirely on computer simulations for its description and analysis, cellular automaton, developed by Von Neumann and Burks (1966). Their application to John H Conway's *game of life* (Gardner, 1970) attracted the inter-

est of both researchers and the general public, and it launched the acceptance of simulations for exploration and confirmation purposes.

Nowadays computers have enabled the study of a large number of complex network structures, which are able to match empirical data and the the arrangements of real social structures. Two main traits have been found to be central in empirical networks in this study. First, most social networks exhibit the so-called small-world property: a combination of short path length (any two individuals can be connected via a small number of connections) and a high clustering (the sets of acquaintances of two friends tend to overlap). These were first replicated with the small-world algorithm of network formation introduced by Watts and Strogatz (1998) in their seminal paper. In this algorithm the links of a regular network are randomly rewired, to connect a different pair of nodes.

Second, virtual networks (such as the World Wide Web) are scale-free: there are very few hubs that connect many scarcely connected nodes. Scale-free networks can be reproduced with the preferential attachment algorithm developed by Barabasi and Albert (1999). New nodes are dynamically introduced to the network, and connected to the already existing nodes with a preference for the ones with a higher number of connections.

These two properties (small world and scale free) are found in most of the empirical network structures in different combinations. They can be merged to reproduce most networks, from scientific collaboration (Newman, 2001) to protein-protein interactions (Bosque et al., 2014) or online social communities (Ahn et al., 2007).

Knowledge production networks have been studied in depth empirically (Balland et al., 2012; Barabasi et al., 2002; Boschma and ter Wal, 2007; Burt, 1992) by exploring how collaboration shapes the creation of novelty. Interacting agents that create knowledge have also been modeled with a game-theoretical approach (Cowan et al., 2007; Konig et al., 2011) and with simulation models (Ahrweiler

et al., 2004), assuming that agents can choose their collaborations to produce new knowledge as a recombination of their joint knowledge pool. Moreover, some attempts have even been made to explore the simultaneous evolution of the network and the knowledge they create, both empirically (Borner et al., 2004) and theoretically (Cowan et al., 2004). A last approach was introduced by Silverberg and Verspagen (2005), who analyzed the creation of knowledge in the network of technologies, from a technology to a complementary or related one.

Knowledge diffusion has been thoroughly studied with contagion and epidemiology approaches (Romero et al., 2011). While epidemiology studies of diffusion focus mainly on ways to prevent the spread of an infectious disease, studies on diffusion of information tend to focus on why information spreads, and what can be done to improve its propagation (Centola, 2011). Recently, the focus has moved from the extent of diffusion in the overall network to the effect of single network characteristics such as path length or clustering coefficient (Zeppini and Frenken, 2015).

1.3 Overview of the thesis

In this thesis, we will explore how insights from network science can be applied to the question of how knowledge is produce in social networks as well as how new knowledge diffuses (or not) through social networks. This thesis is based on the compilation of five scientific studies. Each one of the Chapters 2 to 6 is an individual scientific paper, with its own introduction, results, and conclusion; oriented to publication in peer-reviewed scientific journals. They all follow a common thread, namely the study of the social dimensions of knowledge. Chapters 2 to 4 exploit simulations models, which are extensively used in the social sciences to explore social mechanisms and emergent phenomena (Fagiolo and Dosi, 2003; Helbing, 2012), while Chapters 5 and 6 use an empirical approach.

The thesis is structured as follows. Chapter 2 studies the complex self-reinforcing process of knowledge creation in networks. The amount of knowledge that agents create affect the structure of the network, which in turn influences the future amounts of knowledge created in a path dependent evolution. The simulation model shows that collaboration can be a consequence of attractiveness of the most productive agents that can hamper knowledge creation if they attract too many collaborators and become unproductive. Moreover, a myopic partner selection based merely on previous collaboration can also be harmful for the performance of the system.

Chapter 3 explores the effect of social reinforcement in the diffusion of ideas in a word-of-mouth process in a social network. It compares the efficiency of different network structures for several population types, according to their degree of open-mindedness and homophily. In particular, it focuses on the effect of strong ties (i.e. ties between agents with many common friends) in the overall adoption levels in the population. The main result is that simple, self-evident ideas are better diffused through weak long ties, while complex and controversial ideas benefit from the social reinforcement that comes through strong ties. This effect compensates the difficulty of spreading complex ideas in close-minded populations.

Chapter 4 further elaborates the basic percolation model in Chapter 3 by investigating how social reinforcement determines the transition of scientific paradigms, when several alternative theories attempt to become dominant. If every theory reveals a flaw of the existing paradigm, the self-reinforcing growing discontent creates a cumulative social reinforcement for alternative theories to spread. When the population is critically fragmented between different coexisting theories, a new paradigm emerges by absorbing all adopters of the alternative theories.

Chapter 5 builds on the notion of percolation, but in an empirical sense. Here, we analyze the phenomenon of delayed recognition of breakthrough knowledge, that is, the phenomenon that some inventions initially do not diffuse (percolate), but only later become accepted. The failure in the recognition of these radical developments depends both on the social position of their authors, as would be predicted by percolation models of diffusion in social networks (Chapters 3 an 4), but also on their technological characteristics. Following the literature on delayed recognition and sleeping beauties in scientific papers, it identifies sleeping beauties in technologies as highly cited patents that did not receive citations for a long period, and it compares them to the control group of other highly cited patents from the same period. On the one hand, breakthroughs in the most codified technological areas are more susceptible to earlier appreciation. On the other hand, a collaboration between many of authors, as well as their previous experience, also facilitates preventing delayed recognition.

Chapter 6 analyzes the network structure of occupations that characterize the knowledge base of industries through their division of labor configuration. An exploration of its configuration through recent complex network analysis techniques permits to typify both the industries and the occupations spaces, and to relate them with an economic measure such as their productivity. The knowledge base of an economy can be analyzed by depicting it as a two-mode network mapping professions onto industries. Each industry can then be characterized by the professions present in the industry while, reversely, each profession can be characterized by the industries in which it is present. Occupations show a nested structure: there is a set of common ubiquitous occupations and a set of occupations that are specific, which are used mostly by industries that use many other occupations. Moreover, ubiquitous occupations add to an industry's productivity to a greater extent than specific ones, which are usually the most easily outsourced.

Finally, Chapter 7 summarizes the conclusions and research avenues opened from this thesis. In summary, the main topics that this thesis studies are knowledge creation and knowledge diffusion. The main tools for this come from network analysis techniques. Table 1.1 shows an overview of their use across the different chapters.

	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6
Knowledge creation	×			×	×
Knowledge diffusion		×	×	×	
Network analysis	×	×	×		×

Table 1.1: Overview of the chapter topics and methods

CHAPTER 2

The coevolution of endogenous knowledge networks and knowledge creation

This chapter has been produced in collaboration with JM Azagra-Caro. The PhD candidate has been the primary researcher of the work reported in this chapter and has been the main contributor in all stages of research (idea, theory, methodology, analysis, interpretation, writing and presenting).

2.1 Introduction

Simulation models are useful for explaining evolutionary processes. They have been used widely to analyze knowledge networking, the process by which agents interact to create knowledge. Some interesting examples are inventor network models (Cowan et al., 2006), researcher collaboration models (Grebel, 2012), and interfirm research and development (R&D) alliance networks (Ahrweiler et al., 2004). Simulations have been used also to investigate knowledge creation by agents in a network. In those studies, knowledge is an abstract idea (Cowan and Jonard, 2003), or a concrete output of the abstract knowledge, that can be measured by number of scientific papers in the case of researchers (Borner et al., 2004), new products in the case of firms (Malerba et al., 1999), etc.

Many empirical studies have analyzed the creation of knowledge in networks (for a review, see Ozman, 2009, and Phelps et al., 2012). Most empirical studies use measures of the network related to one agent (the ego network) to explain the agent's output but ignore possible feedbacks. Thus, results are mixed and inconclusive. For example, it is unclear how the simplest measure of ego networks, the number of collaborators (or degree), affects the ego's performance.

A first group of studies indicates that knowledge creation is a collaborative process, that is, that the more intensely agents collaborate, the more knowledge they create (Ahuja, 2000; Boschma and ter Wal, 2007; Cassiman and Veugelers, 2006; Cohen et al., 2002; Ibarra, 1993). This kind of coevolution occurs when collaborations provide agents with resources and new information (Ahuja, 2000). A second group of studies suggests that the number of collaborators is not a determinant of agents' innovative performance (Bell, 2005; Dakhli and De Clercq, 2004; Martinez-del Rio and Cespedes-Lorente, 2013; Vega-Jurado et al., 2009). For example, Bell (2005) finds no significant relation between the number of a firm's formal ties and its innovativeness, because institutional ties transmit only relatively well-known information. Finally, a third group finds that collaborating can harm performance (Grimpe and Kaiser, 2010; Laursen and Salter, 2006; McFadyen and Cannella, 2004; Molina-Morales and Martinez-Fernandez, 2009; Woolcock, 1998). McFadyen and Cannella (2004) suggest that the more collaborations an individual is forced to maintain, the less effort can be focused on knowledge creation. Thus, at some point, agents with too many collaborators may start to underperform compared to those with fewer partners.

These different studies consider an exogenous network structure, that is, they take the structure as given and predict its impact on knowledge creation. This is not realistic, mainly because agents choose their collaborators for specific reasons, including reputation or previous performance (Balland et al., 2012; Wagner and Leydesdorff, 2005). Thus, network formation and knowledge creation are en-

dogenous processes. The structure of the network affects the output of agents, which determines the future structure of the network. Previous empirical studies do not account for this feedback, which may explain the ambiguous results they obtain.

Agents in the network choose their partners for creating knowledge, which changes their future behavior and outputs (Baum et al., 2010). However, few studies propose a theoretical modeling of the interaction between the network and the creation of knowledge. An exception is the work by Cowan et al. (2004) which emphasizes the complex nature of knowledge creation through collaboration. In this study we will propose a model that builds on theirs and additionally analyzes the amount of knowledge created in the network.

Recent studies on endogenous networks (Tedeschi et al., 2014; Vitali et al., 2013) consider companies that form strategic alliances, based on the maximization of an objective function. Those theoretical models allow switching between different strategies (e.g., single innovators, collaborative innovators and imitators) to reproduce some empirical regularities and then design strategies and policies to improve technological innovation, but do not deal with the conflicting empirical evidence on the relation between networks and the performance of agents. We will fill this gap by generalizing companies to any type of knowledge producer and technological innovation to any knowledge output, designing a simulation model of an endogenous and evolving network of agents who create knowledge. Our model does not intend to describe and replicate how knowledge producers interact to create knowledge, nor does it assume any kind of rationality for their collaborations. Instead, agents are attracted to one another for collaborations, and they create knowledge without any assumption on their strategies or objectives in doing so. Simulations of the model generate different theoretical scenarios of the coevolution of knowledge networks and knowledge creation.

The main goal of this study is to call attention to the importance of taking into account the feedback between knowledge creation and network formation when studying knowledge creation in networks. The chapter is structured as follows. Section 2.2 presents the coevolution model of knowledge creation and knowledge networks. Section 2.3 presents the results of the simulations. Section 2.4 discusses the different results and concludes the chapter.

2.2 The model

Let us consider a set $S = \{1, \ldots, n\}$ of agents who interact over T periods of time. In each step $t \in \{1, \ldots, T\}$, they form a network and create a certain amount of knowledge $\kappa(i, t)$. The network is represented by its adjacency matrix Ω_t, where $\Omega_t(i, j)$ takes the value 1 if agents i and j collaborate in period t, and 0 otherwise. The degree (or number of collaborators) of agent i in period t is $d_t(i) = \sum_j \Omega_t(i, j)$. This is a basic indicator to measure the size of the ego network, often used in the empirical literature (see Cooke and Wills, 1999, or Bell, 2005).

The network in each period is formed depending on the network and the knowledge created in the previous periods. As the network is generated in each period, links are allowed to break and form over time. Similarly, the amount of knowledge created in each period depends on the amount of knowledge previously generated, and on the structure of the network. As the process develops, agents become heterogeneous in their knowledge endowments and ego networks. A link between two agents can be created in any period even if they did not collaborate in the past. Likewise, two agents can break an existing collaboration if the link is not updated in a later period.

2.2.1 Knowledge creation

In order to get a model as flexible as possible, we consider knowledge in a very abstract way as in Cowan et al. (2004). Thus, agents in our model could be researchers, inventors, firms, universities, or any other type of agent involved in knowledge creation with the possibility to collaborate.

Prior research has shown that the performance of an agent in a knowledge network can be deeply influenced by the structure of the network (De Solla Price, 1965; Guler and Nerkar, 2012). On the one hand, collaborations can provide resources, new information or new ideas (Ahuja, 2000). This is captured by parameter $\theta \geq 0$, the positive effect of collaborations in knowledge creation. On the other hand, collaborations can be costly (McFadyen and Cannella, 2004; Ozman, 2009), with the result that a very large number of collaborations can hamper knowledge creation. Additionally, the cost of maintaining an additional collaboration increases with the number of collaborations an agent already maintains. This is captured by parameter $\gamma \geq 0$, the cost of collaborating, together with the square of the number of collaborations d_t.

Finally, knowledge is a cumulative process: new knowledge can be created from previous knowledge (Jaffe et al., 2000). Parameter α measures how much new knowledge is created from the stock of knowledge of an agent, and hence, accounts for the cumulativeness of knowledge. The length of the time window is τ, which is the number of periods before knowledge becomes obsolete.

Equation 2.1 shows the functional form for the creation of knowledge. The amount of knowledge created by agent i at time t, $\kappa(i,t)$, depends on the structure of its ego network and on the stock of knowledge it possesses. As stated, θ is the positive effect of collaborations, γ is the cost of collaborating and α is the knowledge produced from the stock.

$$\kappa(i,t) = \theta\frac{1}{\tau+1}\sum_{s=t-\tau}^{t} d_s(i) - \gamma d_t(i)^2 + \alpha\frac{1}{\tau}\sum_{s=t-\tau}^{t-1}\kappa(i,s) \tag{2.1}$$

This functional form implicitly assumes that there are only two possible sources of new knowledge for an agent: collaborations and the pool of knowledge the agent already possesses. Due to the recombinant nature of knowledge (Konig et al., 2011, 2012), if an agent never collaborates, the possible number of new combinations of its existing knowledge are limited. At some point, the agent will be unable to continue to create new knowledge unless it starts to collaborate. Given this functional form for the creation of knowledge, α is necessarily bounded to $[0,1]$.

This function is similar to that by Konig et al. (2011, 2012). In their model, the amount of knowledge created by an agent is a recombination of the knowledge stocks of the agent and his neighbors. In line with this approach, we add a new dimension to the effect that the knowledge of neighbors has on the knowledge of an agent. In our model, the knowledge stocks of potential collaborators influence the amount of knowledge agents create, through their influence on whether or not they become collaborators. Although it does not appear explicitly in the function of knowledge creation, it implicitly affects the result.

2.2.2 Network formation

In a knowledge network, agents break and establish links as a result of strategic decisions (Barabasi et al., 2002; Fleming and Frenken, 2007). These strategic decisions usually depend on two main components, previous history and attractiveness of agents (Ahrweiler et al., 2004). More experienced and more successful agents are more likely to find partners (Balland et al., 2012; Wagner and Leydesdorff, 2005), so the probability of collaborating increases with the attractiveness of the agent. In our model, the attractiveness of an agent depends on the amount

of knowledge it created previously, relative to the knowledge created by the other agents in the network. Moreover, previous collaboration increases the willingness to engage in joint knowledge creation (Baum et al., 2010; Cowan et al., 2006), so the probability that a link is formed is higher for pairs that have already collaborated.

We model the probability that agent i collaborates with agent j as a linear combination of their previous history and the attractiveness of agent j (Equation 2.2), respectively weighted by λ and $1-\lambda \in [0,1]$. The previous history of a couple of agents is the number of times they have collaborated in the recent past, during the time window. The attractiveness of an agent is the knowledge it has created in a period as a proportion of the maximum amount of knowledge created by itself or another agent in that period. If the agent is the maximum knowledge producer, this proportion will be equal to 1. This function follows a preferential attachment dynamic (Albert and Barabasi, 2002; Barabasi and Albert, 1999), as agents with a high number of collaborators are likely to attract even more collaborators.

$$P(i \to j, t) = \lambda \frac{1}{\tau} \sum_{s=t-\tau}^{t-1} \Omega_s(i,j) + (1-\lambda) \frac{1}{\tau} \sum_{s=t-\tau}^{t-1} \frac{\kappa(j,s)}{\underset{k}{max}\ \kappa(k,s)} \tag{2.2}$$

The probability that agents i and j collaborate is that both of them collaborate with the other (Equation 2.3). In considering probabilities we account for their being willing to collaborate but unable to do so for some reason. The probability that a link breaks, that is, that a collaboration in time $t-1$ does not continue in time t, is $1 - P(i \leftrightarrow j, t | \Omega_{t-1}(i,j) = 1)$, the probability that it doesn't form in t given that it existed in $t-1$. The probability that it breaks after $s \in \{1, .., \tau\}$ periods is defined likewise. A summary of the parameters can be found in Table 2.1.

$$P(i \leftrightarrow j, t) = P(i \to j \cap j \to i, t) \tag{2.3}$$

Param.	Interpretation	Constraints
θ	Positive effect of collaborations in knowledge creation	$\theta \geq 0$
γ	Cost of collaborating	$\gamma \geq 0$
α	Amount of knowledge created from the stock	$0 \leq \alpha < 1$
λ	Weight of previous collaboration in the probability to collaborate	$0 \leq \lambda \leq 1$
τ	Length of the time window	$\tau \in \mathbb{N}$

Table 2.1: Parameters of the model

Notice that agents' actions are not the result of an explicit decision making process since they do not develop strategies that maximize some objective function, or compete over knowledge, as in the theoretical literature (see e.g. Azagra-Caro et al., 2008 or Westbrock, 2010) or previous work on the coevolution of knowledge and networks (Cowan et al., 2004; Llerena and Ozman, 2013). In this chapter, we model the probability of a collaboration between two agents, not the rationale for their collaboration which is beyond the objectives of this study.

2.3 Simulation results

In this section we present different scenarios for several sets of parameters. For the simulations, we consider a set of $n = 200$ agents interacting for $T = 500$ periods. The results of the simulations are not qualitatively different depending on the length of the time window τ. In the simulations we use a value of $\tau = 1$, although the results are similar for other window lengths. Every agent starts with 1 unit of knowledge. Every pair of agents is connected with probability $2/n$, which results in a moderately dense initial network. Similar results were obtained with other initial conditions, like binomial knowledge endowments or other probabilities for the initial network. The initial conditions were the same in all the settings but resulted in very different final configurations. At the beginning of the simulation, agents are homogeneous in their knowledge endowment and heterogeneous in their ego network, and as the process unfolds, they become

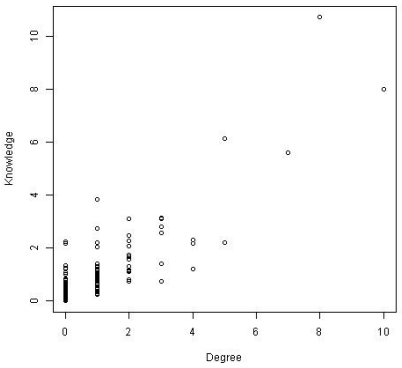
heterogeneous in both the knowledge they have created and the network they are forming.

2.3.1 Positive, negative and independent coevolution

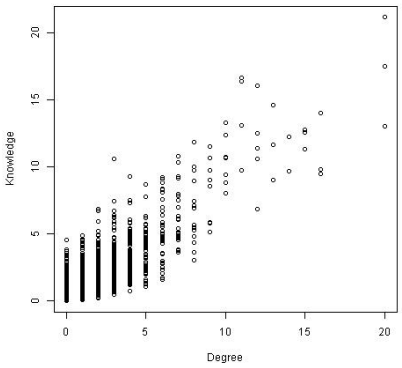
The behavior of the model is depicted graphically. Figures 2.1, 2.2 and 2.3 are examples of the three possible scenarios described in the empirical literature.

When the positive effect of collaborations, θ, is high enough compared to the cost of maintaining a collaboration, γ, collaborations are profitable. This results in a positive relation between the number of collaborations and the creation of knowledge (Figure 2.1). Figure 2.1a shows an example of a reinforcing effect in knowledge creation for the whole population. Most agents create moderate amounts of knowledge, with a low number of collaborations. A few of them, nonetheless, are more productive and attract more collaborators, and thus produce even more and are more attractive in the following steps.

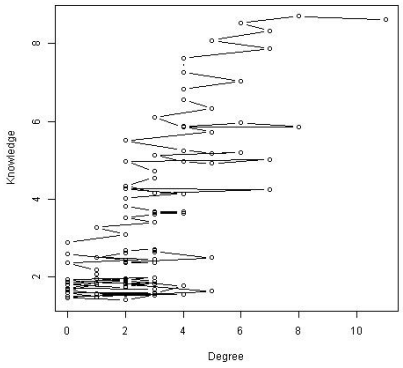
As a robustness check, we run 20 simulations with the same parameter settings and plot the last period of all simulations. Since we overlap the points of all runs, spurious relations between the different scatter plots could appear (for instance, if two runs gave identical independent plots, but one of them higher and to the right of the other, the overlapped plot would look like a positive plot). Thus, the robustness check is presented as an additional plot (Figure 2.1b), in addition to the snapshots of single simulations. Figure 2.1b shows the last period of the simulation for 20 simulations with the same parameters as Figure 2.1a. Both figures show a positive relation between the degree of agents and the knowledge they produce, even though some of the agents in some of the runs in Figure 2.1b reached higher amounts of knowledge and degree. This shows that the aggregated coevolution pattern depends only on the parameters of the model, not on the initial conditions. On the other hand, the particular amounts of knowledge



(a) Whole population, final outcome (one run).
$\alpha = 0.9$, $\theta = 0.1$, $\gamma = 0.0001$, $\lambda = 0.2$



(b) Whole population, final outcome (20 runs).
$\alpha = 0.9$, $\theta = 0.1$, $\gamma = 0.0001$, $\lambda = 0.2$

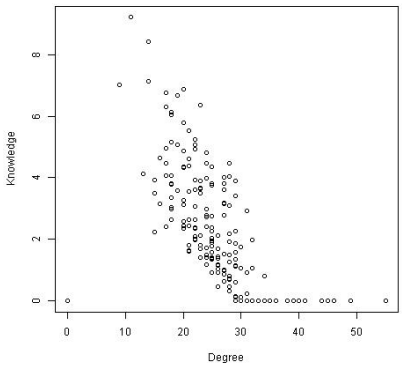


(c) A single agent, dynamics (one run).
$\alpha = 0.9$, $\theta = 0.1$, $\gamma = 0.0001$, $\lambda = 0.2$

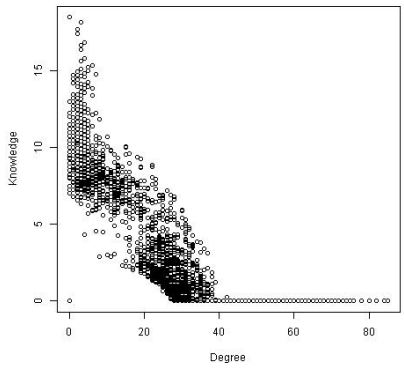
Figure 2.1: Positive coevolution

and degree do depend on how the simulation develops, that is to say, on the probabilities involved in the network function (Equation 2.3).

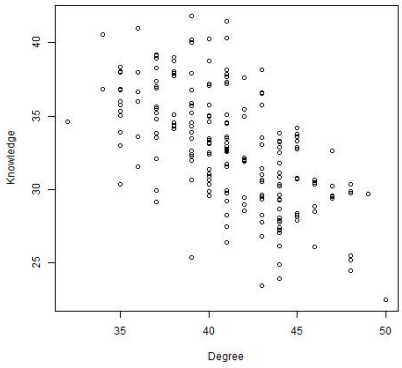
Figures 2.1*a* and 2.1*b* show the results for the whole population in a single period, the last one. The dynamics for a single agent are depicted in Figure 2.1*c*: the more knowledge it creates, the more collaborators it gets, and the more collaborators it has, the more knowledge it creates. This leads to the differentiation between those agents with low levels of both knowledge and collaborations, and those that attract most collaborations and create higher amounts of knowledge.



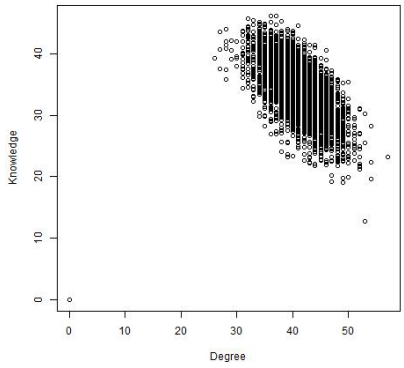
(a) Whole population, final outcome (one run).
$\alpha = 0.1, \theta = 0.5, \gamma = 0.01, \lambda = 0.2$



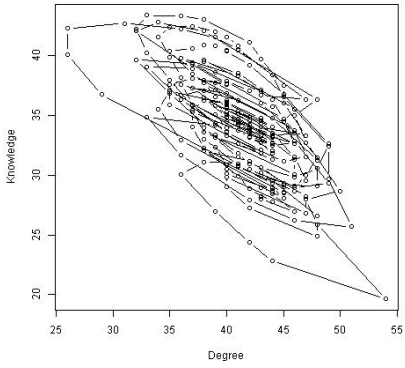
(b) Whole population, final outcome (20 runs).
$\alpha = 0.1, \theta = 0.5, \gamma = 0.01, \lambda = 0.2$



(c) Whole population, final outcome (one run).
$\alpha = 0.9, \theta = 0.5, \gamma = 0.01, \lambda = 0.8$



(d) Whole population, final outcome (20 runs).
$\alpha = 0.9, \theta = 0.5, \gamma = 0.01, \lambda = 0.8$



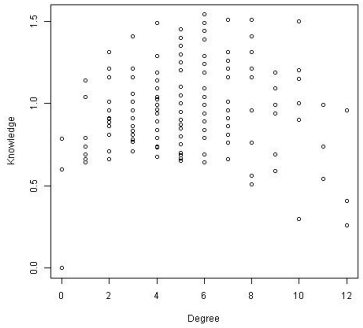
(e) A single agent, dynamics (one run).
$\alpha = 0.9, \theta = 0.5, \gamma = 0.01, \lambda = 0.8$

Figure 2.2: Negative coevolution

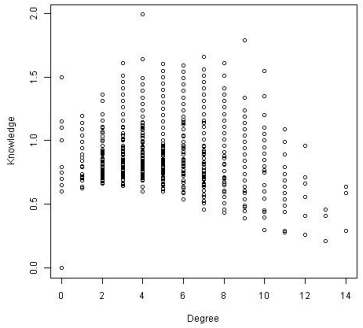
If the cost of maintaining collaborations, γ, is high compared to their positive effect, θ, every additional collaboration is prejudicial to the agent's performance. In this case, the amount of knowledge created is decreasing in the number of collaborations, and the observed coevolution is negative (Figure 2.2). The existence of a limitation to collaboration is a necessary condition for this coevolution pattern. In some cases, the cost of collaborating can be so high that agents have too many collaborations to produce any knowledge (Figures 2.2*a* and 2.2*b*). In other cases, the relation can be negative but the amounts of knowledge created can still be high enough to avoid the appearance of unproductive but collaborative agents (Figures 2.2*c* and 2.2*d*). In negative coevolution scenarios, the most productive agents become the most attractive. As their number of collaborations increases, they create less and less new knowledge and become less and less attractive. At some point, they have a smaller number of collaborations which allows them to perform better, and thus to become more attractive again. These dynamics are depicted in Figure 2.2*e* for a single agent.

Finally, the amount of knowledge created through collaborations may be similar to the amount based on previous knowledge. Due to the mixed effect of both sources of new knowledge, the performance of agents with many collaborators, and agents with small numbers of collaborators but a large stock of knowledge, is similar. Thus, the coevolution is independent, since the number of collaborators seems not to affect the creation of knowledge (Figure 2.3). This pattern can appear both for low values of degree and knowledge (Figures 2.3*a* and 2.3*b*) and for higher values (Figures 2.3*c* and 2.3*d*). This last case shows a parameter setting in which agents can have a high number of collaborations.[1]

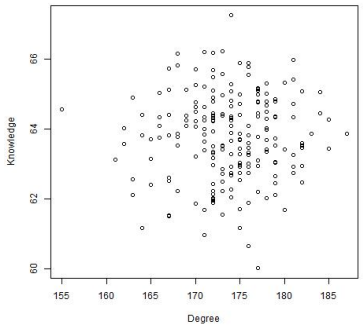
[1]Notice that our agents can be any kind of knowledge producers. In certain academic disciplines, for instance, very high numbers of collaborators are plausible: the original human genome paper (doi:10.1038/35057062) has 249 authors, and the Higgs Bosson paper (doi:10.1016/j.physletb.2012.08.021) is signed by 2891 authors.



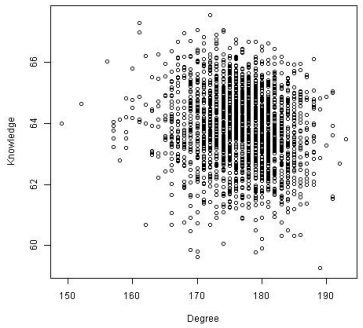
(a) Whole population, final outcome (one run).
$\alpha = 0.9, \theta = 0.1, \gamma = 0.01, \lambda = 0.5$



(b) Whole population, final outcome (20 runs).
$\alpha = 0.9, \theta = 0.1, \gamma = 0.01, \lambda = 0.5$



(c) Whole population, final outcome (one run).
$\alpha = 0.1, \theta = 0.5, \gamma = 0.001, \lambda = 0.2$

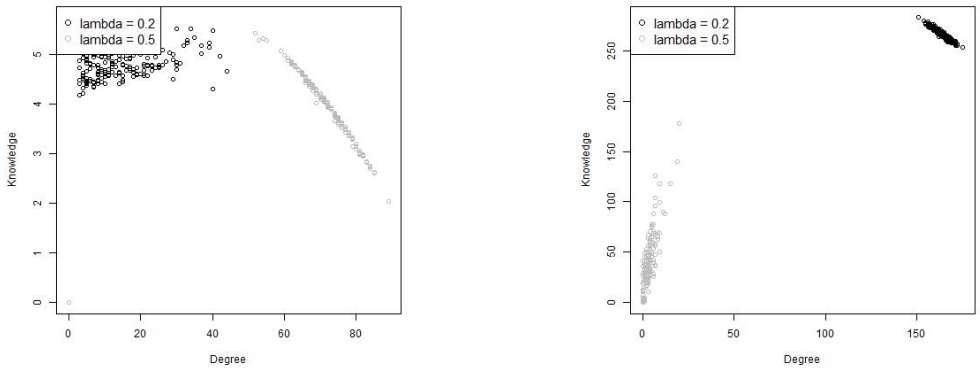


(d) Whole population, final outcome (20 runs).
$\alpha = 0.1, \theta = 0.5, \gamma = 0.001, \lambda = 0.2$

Figure 2.3: Independent coevolution

2.3.2 Feedback mechanisms in the model

In this section we show the determinant role of the feedback between the two functions. Some variations in a single parameter are illustrative of the complexity of the model described. The straightforward effects of the model are that changing parameters in Equation 2.1 varies the amount of knowledge created, and that changing parameters in Equation 2.2 varies the network structure. Nonetheless, the model also shows the indirect effects of each equation in the outcome of the other, due to the relation between the two functions. Figures 2.4 and 2.5 are examples of the complex behavior of the model due to the feedback between the processes of knowledge creation and network formation.

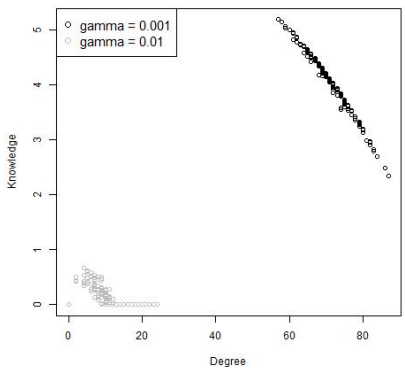


(a) Whole population, final outcome. $\alpha = 0.5, \theta = 0.1, \gamma = 0.001$

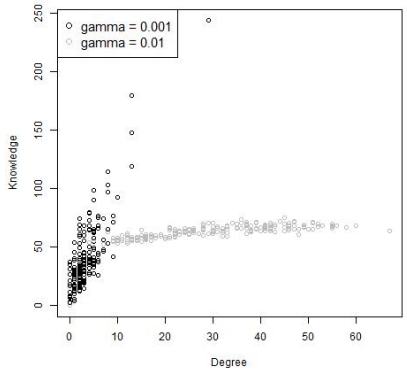
(b) Whole population, final outcome. $\alpha = 0.9, \theta = 0.9, \gamma = 0.045$

Figure 2.4: Effect of an increase in λ

Consider that we fix all the parameters in the knowledge creation function and vary only λ, the weight of having previously collaborated on the probability to collaborate, from 0.2 to 0.5 (Figure 2.4). When $\lambda = 0.2$, the probability to collaborate depends mainly on the attractiveness of agents; while when $\lambda = 0.5$, it depends also on whether or not the two agents have collaborated in the previous steps. With this increase in λ, the process can change the type of coevolution case. Figure 2.4*a* shows a switch from an independent to a negative coevolution scenario, and Figure 2.4*b* depicts a swap from a negative to a positive coevolution. In both cases, the apparent relation between the number of collaborators and the amount of knowledge is not driven by the knowledge creation function but by the network formation process. In Figure 2.4*a*, with a high value of λ, only the most productive agents attract new collaborators. This leads to a dynamic similar to the one previously depicted in Figure 2.2*e*. Likewise, in Figure 2.4*b* the relation is positive not because agents with more collaborators create more knowledge, but because the most productive agents attract more collaborators. In this case, the positive coevolution scenario performs worse than the negative coevolution case. When agents can choose their collaborations freely, based on attractiveness rather than previous history, the overall amount of knowledge created is higher.



(a) Whole population, final outcome.
$\alpha = 0.5$, $\theta = 0.1$, $\lambda = 0.5$



(b) Whole population, final outcome.
$\alpha = 0.9$, $\theta = 0.9$, $\lambda = 0.2$

Figure 2.5: Effect of an increase in γ

A similar behavior can appear for changes in parameters of the knowledge function as well. Figure 2.5 shows two examples of increasing γ, the cost of collaborating, from 0.001 to 0.01. Of course, the amount of knowledge created will be higher for lower values of this cost. Moreover, increasing γ can lead to a change in the structure of the resulting knowledge network. Figure 2.5*a* shows a case where the cost of collaborating is so high that it suffocates the whole process: the amount of knowledge created is lower, and the network does not develop. Figure 2.5*b*, on the other hand, shows a switch from a positive to an independent coevolution. In the positive coevolution, some agents have the role of "star scientists": they create a much higher amount of knowledge than the rest, and they are more attractive to collaborators. With the increase in γ, the reinforcement mechanism disappears and all agents create similar amounts of knowledge. They are also similarly attractive, so the degrees are more uniformly distributed in the resulting network.

2.4 Discussion and conclusion

This chapter presents a simulation model of the coevolution of knowledge networks and knowledge creation. Knowledge appears increasingly to be an interactive process, and a deeper understanding is needed to improve the efficiency of the performance of the system. Different coevolution scenarios can arise depending on the importance of the collaborations for knowledge creation, the role of previous knowledge, and the process of partner selection. Two rules of behavior reproduce the scenarios in the empirical literature, which correspond to different cases of apparently conflicting empirical evidence. Thus, all those different cases of knowledge creation through collaboration originate in a single process.

The main result of the chapter is the importance of feedback between the knowledge creation process and the formation of the network. Modifying the partner selection process from depending mainly on attractiveness to depending also on whether two agents have already collaborated leads to changes not only in the structure of the network, but also in the amount of knowledge agents produce overall. It can also lead to a change of coevolution type (from negative to positive, for example, or from independent to negative). Likewise, changes in the knowledge creation function, like a variation in the cost of collaborations, produces changes not only on the amount of knowledge created overall, but also on the type of coevolution and the structure of the resulting network. These feedback mechanisms must therefore be accounted for in studies aiming to analyze a process of knowledge creation in a network.

The results draw attention to several other features of the process. First, the positive relation between the number of collaborators and the performance of agents may be due to partner selection rather than knowledge creation. This has implications for researchers since it points to the importance of taking into ac-

count the endogeneity of the network when analyzing the effect of collaborations in knowledge creation.

The process of knowledge creation can be hampered by "myopic" partner selection based on previous history rather than attractiveness. If agents are bound to their previous collaborations, the overall performance of the system is lower; if they can freely establish new links or break existing ones, the levels of performance achieved by the system are higher. If previous history is important because of high levels of uncertainty and instability, one solution might be to improve the legal framework in order to reduce the risk of hold-up, and thus increase the willingness to interact with unknown partners. In the case of a social context where agents have few opportunities to meet new partners and start new collaborations, encouraging agents to increase the number of their collaborations might be enough to force the creation of linkages with new partners. When the goal is to increase both knowledge creation and collaboration, this can be achieved by focusing on improving collaboration.

Despite the general belief that collaborations boost performance and knowledge creation, it is important to remember that collaborations are not costless. The cost of establishing and maintaining a collaboration can sometimes outweigh the benefits of collaborating. In a dynamic setting, it might switch the relation between knowledge creation and networking, so that the most productive agents may attract too many collaborators and become unproductive.

This chapter has some limitations. First, the simulation model suggests different lines of action for different underlying processes. Selecting the right process is essential for choosing the right action to implement. In order to address this, future research would benefit from empirically validating the model. This could be done with the Werker and Brenner (2004) method, or by calibrating the parameter values with part of a panel of patent data and then checking how this calibrated model fits the rest of the data from the panel. This empirical validation

may allow comparison of different knowledge creation processes. Furthermore, the model could be adapted to incorporate parameter changes through time in order to implement policy actions. Then, the policy actions suggested for the different scenarios could be tested through simulations, in a secure and costless way. However, measuring the parameters needed for every application with sufficient precision would be difficult, even more so since each parameter value is likely to be heterogeneous for each actor. This either increases the amount of data needed to calibrate the model to fit a specific problem or reduce its predictive power, if parameter value at the population level are left constant. How to solve these issues could be subject of future work.

Nonetheless, the main goal of this chapter was to call attention to the importance of the feedback between network formation and knowledge creation. This has been accomplished by showing that changes in one of the two functions (the knowledge or the network) can affect the results of the other (the network or the knowledge). Taking into account the endogeneity of the process is crucial when studying knowledge creation in networks. Scholars aiming to improve the empirical literature should tackle this issue and not take the network as fixed or given to study its (on-way) effect on the performance of its agents.

CHAPTER 3

Percolation with social reinforcement

This chapter has been produced in collaboration with K Frenken and P Zeppini. The PhD candidate has been the primary researcher of the work reported in this chapter and has been the main contributor in all stages of research (idea, theory, methodology, analysis, interpretation, writing and presenting).

3.1 Introduction

Ideas spread often through social contact, and the structure of societies plays a primary role in the diffusion process. This is even more true in recent years, due to the advent of digital social networks like Faceboook and Twitter. Although many studies have analyzed how the social structure affects the diffusion of information, not much is known about the role of social influence in and heterogeneity in the diffusion process in a network.

The current access to massive online databases of agents interacting offers a never-failing source of empirical data of information diffusion. One can observe how agents share, like or retweet pages, photos or messages. This has lead to a new strand of literature analyzing the behavior of information diffusion in such

online social networks (Dow et al., 2013; Lehmann et al., 2012; Romero et al., 2011; Ugander et al., 2012; Wu et al., 2011).

In the present chapter we propose a model of diffusion in small-world networks addressing the interplay of network structure, social influence, and heterogeneity of agent's characteristics. We show that the so-called "weak" ties of small-worlds promote the diffusion process in the absence of social pressure in the neighborhood of an agent, while the "strong" ties of close-knit cluster are more important when social pressure is present. This is specially important for populations with varying degrees of openness to new ideas.

Our model builds upon a percolation framework (Zeppini and Frenken, 2015) in order to have a clear benchmark for diffusion processes on networks that rely exclusively on individual adoption decisions (without social pressure) and we study the interplay of individual openness and social reinforcement, so as to understand the role of structural factors such as clustering as well as the role of openness distributions. Moreover, we assume that ideas diffuse by mean of word-of-mouth communication on friendship contacts (Alkemade and Castaldi, 2005; Campbell, 2013). This work has two main focuses: first, we systematically study the effect of local social reinforcement in a diffusion process, and classify network structures in terms of their efficiency for different degrees of clustering. Second, we argue that not only the structure of the network matters for diffusion efficiency, but also the distribution of individual agents openness to new ideas.

We specify agents' heterogeneous characteristics in terms of their "openness" to a new idea, and find that with a uniform distribution of openness fully random networks are more efficient than small-worlds, both with and without social reinforcement. This result goes against the results of Centola and Macy (2007). On other hand, for less open populations the social reinforcement mechanism in close-knit clusters is more important, and small-worlds are more efficient than fully random networks.

The structure of the chapter is as follows. Section 3.2 is a selection of relevant literature on diffusion processes. Section 3.3 presents the basic percolation model. The model is extended with social reinforcement (Section 3.4), homophily (Section 3.5) and non-uniform distributions of openness (Section 3.6). Finally, in Section 3.7 we present some conclusions.

3.2 Literature review

Understanding how information is diffused in a social network has many implications, from the spread of vaccination in a population to the use for social media for marketing purposes. The analysis of diffusion in a network has been tackled through many approaches over the years.

3.2.1 Models of diffusion in networks: epidemiology

Models of diffusion first started as models of infectious diseases in the field of epidemiology. One of the first models analyzed malaria, introducing a vector (the mosquitoes) that transmitted the disease between humans in a set of differential equations dependent on the birth and biting rates of mosquitoes (Ross, 1915). Current epidemiology models consider direct transmission, from an infected individual to a susceptible one. Transmission from an infected individual to several others is usually well understood, but mathematical models help to comprehend the large scale dynamics of the spread of an infectious disease in a population (Hethcote, 1989).

The seminal models of contagion are the so-called SIR (Kermack and McKendrick, 1927). The population is divided in three exclusive classes: the susceptible (S), the infected (I), and the removed (R). The susceptible fraction of the population may be affected by the disease, the infected fraction can transmit the disease to the susceptible population, and the removed fraction have either

recovered from the disease and gained immunity, or have died and thus are removed from the original population. As they are fraction of the overall population, $I + S + R = 1$.

In a SIR (susceptible-infected-recovered) model, susceptible individuals can be infected by a contagious peer, and after a period of illness they recover or are removed (Kermack and McKendrick, 1927). These SIR models are suitable for diseases produced by viruses, since the infected agents can gain immunity to infection after their recovery. A second family of models are the so-called SIS (susceptible-infected-susceptible). In such models, infected individuals do not gain immunity after recovery, so they become susceptible to contagion again. These SIS are suitable to model, for instance, bacterial infections such as the plague or venereal diseases. Other combinations are also possible, such as SI (no recovery), SIRS (temporary immunity), etc. Also, other intermediate states can be defined, such as an incubation state E when the infected agent is not yet contagious, or a group P born with a passive immunity acquired from the mother.

The basic models consider that the population is constant and all individuals have the same probability of contagion after contact with an infected individual and the same probability of recovery. It can be extended with birth and death rates, differences in the resistance to contagion, differences in the speed of recovery, etc. Later developments include transmission in a social network: susceptible individuals can only be infected by infected individuals in their social ego network, like friends or neighbors (Kuperman, 2013). These models have been applied to several network topologies such as small-worlds (Moore and Newman, 2000), scale-free networks (Pastor-Satorras and Vespignani, 2001), or adaptive evolving networks (Zanette and Risau-Gusman, 2008). In these studies, the focus is on which kind of network is most resistant to the spread of a disease. Small world networks can vary from an endemic state, to periodic oscillations in the

number of infected individuals, depending on the rewiring parameter (Kuperman and Abramson, 2001). In scale-free networks, a disease will always propagate, and it will be concentrated in the individuals with the highest degree (Newman, 2002). Studies with evolving networks have shown that isolation of infected individuals can be an effective control strategy (Risau-Gusman and Zanette, 2009).

3.2.2 Information diffusion

There is a wide and growing body of literature studying the diffusion of ideas in a network. Several approaches can be used to model information diffusion in social networks. One particular case of information diffusion in networks can be found in the opinion dynamics literature. Studies on opinion dynamics analyze the dynamics of agreement in a population. This has lead to models of voters during elections, in which the opinion of an individual can be represented as a number in the interval $[0, 1]$, being 0 and 1 two opposed extremist opinions. The opinion of an individual can be influenced by other people in their social networks. Individuals with extremist opinions can be more vocal about their opinion, and also more difficult to convince to change their opinion. Moreover, social networks tend to be homophilious: individuals are more likely to be friends with people with a similar opinion to theirs.

The spread of information in a social network can also be analyzed with epidemiological models. Databases of online social networks such as Twitter and Facebook have recently been the center of several studies of information diffusion (Dow et al., 2013; Lehmann et al., 2012; Romero et al., 2011; Ugander et al., 2012; Wu et al., 2011). Many of them notice a correlation between the number of friends engaging in a behavior, such as sharing a page or a tweet, and the probability of adopting the behavior (Bakshy et al., 2012). Such processes of information diffusion are what (Centola et al., 2007) called a complex propagation, as opposed to a simple propagation where they are not correlated. In

a simple propagation one single active friend is enough to trigger adoption, the following adopting friends are redundant. A recent study by Romero et al. (2011) have found that both kinds of information coexist in Twitter. Twitter idioms (i.e., #ilikeitwhen) behave as a simple propagation: a single exposure to an idiom is usually enough for users to decide whether they adopt it. The adoption of politically controversial hashtags, on the other hand, is specially affected by multiple repeated exposures (Romero et al., 2011).

Simple and complex propagations interact differently with the network structure. Simple propagations are diffused through weak ties, that connect otherwise disconnected components of the network, so they are better diffused through open networks like random networks. Complex propagations, on the other hand, are diffused better through strong ties that connect individuals with many common friends: if a friend B of an individual A adopts, it is likely that some of the shared friends of A and B will adopt as well, increasing the likelihood of adoption for A. Thus, complex propagations diffuse better in networks with many strong ties, like ring lattices or small worlds.

3.2.3 Models of diffusion in networks: percolation

Epidemic models of diffusion in networks are extensions, introducing a network, to models initially focused in something else. Nonetheless, some models have been developed in physics to analyze diffusion in networks. These models allow for agent diversity and different network structures, and can be extended to include different cases. One such model is percolation (Solomon et al., 2000). Percolation models were originally developed in physics to study how liquids are filtered through porous materials, depending on the internal structure of the material, the composition of the liquid, etc. These models can be used to analyze how innovations or ideas are filtered through the fabric of society, across a social structure.

Percolation models can be used to analyze a word-of-mouth process of diffusion. The main feature of a percolation process of diffusion is that all information comes determinstacally through the social network, by adopting contacts, and only at the moment of adoption. This is different from epidemic models of diffusion, in which a contagion has positive probability after contact, dependent on the resistance of the individual. It is also different from other models of information diffusion where information can come both from the social network and from external sources such as the news. As such, percolation is suitable for the propagation of rumors through word-of-mouth. It is also adequate to model the diffusion of messages through Twitter, since all messages that a user receives come through their contacts, and users receive messages only once, at the moment of publication.

3.3 Basic percolation model

We analyze the diffusion process of new ideas on a population of potential adopters that are embedded in a social network structure. Ideas are identified by their value, represented by a number $v \in [0, 1]$. Agents are heterogeneous in their openness, or resistance to adopt the idea. In an epidemic model, this openness would be equivalent to the resistance to contagion. They are characterized by their minimum quality requirement (*MQR*) for adopting a new idea. The higher the *MQR* - the more closed an agent is - the higher the value he requires of an idea in order to adopt it. In this section, the *MQR* of agents is a random variable which is uniformly distributed, $q \sim U[0, 1]$. In the next section we will consider non-uniform distributions.

The theoretical framework just presented corresponds to the so-called social percolation model (Solomon et al., 2000; Zeppini and Frenken, 2015). In this

framework time is discrete, and agents adopt the new idea at any given time t if the following three conditions are met:

- the agent has not adopted before t,
- the agent is informed, which only occurs if at least one neighbor has adopted at time $t-1$,
- the value of the idea is higher than the *MQR* of the agent, that is $q_i < v$. We name those agents 'willing-to-adopt'.

In a well-mixed population, without network structure of social contacts, there is perfect information. As soon as the idea enters the society, the willing-to-adopt agents adopt, while the rest do not. Since the *MQR* is uniformly distributed as $q_i \sim U[0,1]$, a proportion $100 \cdot v\%$ of the population will adopt an idea of value $v \in [0,1]$. This case can be represented in our model with a complete network, where every agent is connected to every other agent. In a complete network, a single early adopter will inform the whole population of agents about the existence of the idea. When agents are embedded in a social network structure instead, and information travels only through social contacts, two different regimes in the *MQR* space v arise: a *diffusion* regime, where the diffusion size is about the same that one obtains in a well-mixed population, and a *no-diffusion* regime, where diffusion is almost absent. These two regions are separated by a percolation *threshold* v_c, as the result of a second-order *critical transition* (Stauffer and Aharony, 1994).

The percolation process just described is an instance of *simple* propagation. Zeppini and Frenken (2015) present a systematic analysis of networks efficiency and the effect of clustering in this diffusion process.

3.3.1 Network structure

In a percolation setting, agents become informed of the existence of the idea through their neighbors. Thus, the structure of the social network where the agents are embedded can be a determinant of the outcome of the process. Previous studies have considered percolation processes in regular networks as a two dimensional lattice (Cantono and Silverberg, 2009; Hohnisch et al., 2008; Zheng et al., 2013) or a completely random network (Campbell, 2013).

The clustering coefficient of a network is the relative number of triads present in the network. In clustered networks, the probability of two agents being connected increases if they have a shared neighbor. Although their simplicity can be useful for their implementation and the interpretation of the results, a comparison between square lattices and random networks does not allow for an analysis of the effect of clustering in the diffusion process.

In this chapter we propose the use of the small world algorithm (Watts and Strogatz, 1998) for the modeling of the social structure as in Cowan and Jonard (2004). This provides with a family of networks, an interpolation between regular lattices and completely random networks. The algorithm starts with a regular ring lattice and rewires every link with probability μ.[1] This parameter allows to fine tune the randomness of the network.

Varying the rewiring probability μ of the small world algorithm produces networks with varying average path length and clustering coefficient (Figure 3.1). The case with $\mu = 0$ is the one-dimensional regular lattice, and the case with $\mu = 1$ is the random network, also known as Poisson network or Erdos-Renyi model (Erdos and Renyi, 1959). For intermediate probabilities μ, the resulting networks present intermediate clustering coefficients. The "typical" small world

[1] An alternative algorithm proposes to add links instead of rewiring the existing links (Newman and Watts, 1999). This alternative, nonetheless, changes the density of the networks. To ease the comparison between networks, we use the original algorithm.

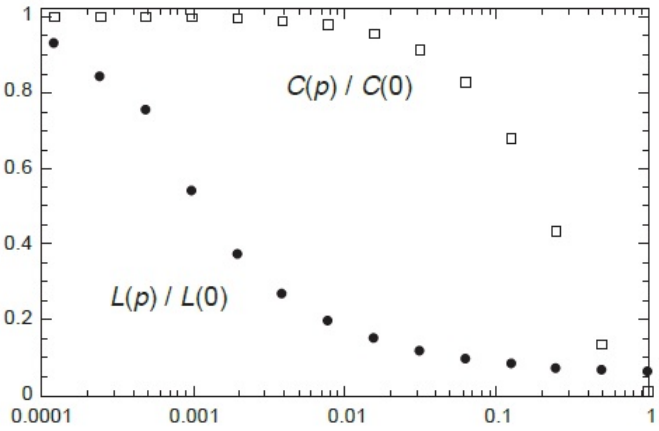


Figure 3.1: Clustering coefficient C (white squares) and average path-length L (black dots) as a function of the rewiring probability in small world networks (Source: Watts and Strogatz, 1998)

is the one with rewiring probability $\mu = 0.01$, presenting an average path-length almost as low as the Poisson network, while still having a clustering coefficient which is comparable with the one-dimensional regular lattice.

3.3.2 Simulation of the percolation benchmark

We analyze the percolation model by means of batch simulation experiments. Figure 3.2 compares the final percentage of adopters in several network structures from the small world model (Watts and Strogatz, 1998), with rewiring probability $\mu \in \{0, 0.001, 0.01, 0.1, 1\}$. All networks have $N = 10{,}000$ nodes, each one representing a potential adopter of the idea, with $k = 4$ neighbors on average. The *MQRs* of agents are random draws from a uniform distribution, $q^i \sim U[0, 1]$. In order to minimize the effect of "'lucky draws", the results show the mean percentage of adopters over $R = 50$ runs of the same scenario, or combination of network structure and value of the idea $v_0 \in [0, 1]$. In all simulations the diffusion process is initialized with 10 randomly chosen early adopters, the seeds of the simulation.

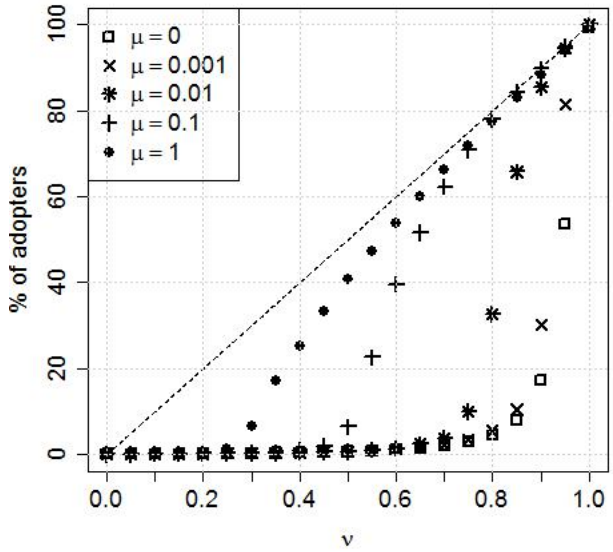


Figure 3.2: Diffusion size in different small world networks for different initial values $v \in [0,1]$ of the diffusing idea (horizontal axis) for different network structures. Reported values are averages over 50 simulation runs. The network size is $N = 10{,}000$ nodes, with 10 seeds.

The dashed line shows the cumulative distribution of the uniformly distributed *MQR* values. It shows the percentage of nodes that are willing to adopt an idea of value v (Equation 3.1).

$$q \sim U[0,1] \text{ and } v \in [0,1] \quad \longrightarrow \quad P(q \leq v) = v \tag{3.1}$$

If every node was informed, the percentage of adopters would follow this dashed line. Nonetheless, the social structure creates "information failures" compared to the well-mixed population with perfect information. Some willing to adopt agents never become informed of the existence of the idea because none of their neighbors have adopted it. Thus, the final diffusion size is lower than the linear demand (dashed line).

The diffusion sizes in Figure 3.2 show two different phases: a non-diffusion regime and a diffusion regime. The phase change is produced by a sharp increase

at the percolation threshold. The thresholds are approximately 0.3 for the random network, 0.8 for the ring lattice, and 0.5, 0.7 and 0.8 for the intermediate rewiring probabilities 0.1, 0.001 and 0.001, respectively.[2]

In the benchmark percolation model, clustering hampers diffusion. Figure 3.2 shows that percolation thresholds decrease with μ, while diffusion sizes increase with μ. Low values of rewiring probabilities μ produce networks with a large short path length and a high clustering coefficient. In such networks, most links are redundant and cannot be used to reach new sources of information. In this simple propagation case, information travels faster through random networks ($\mu = 1$).

3.4 Social reinforcement

We extend the model above to *complex* propagation processes by introducing local social reinforcement. While in the model of *simple* propagation only the first time that an agent is informed about the new idea determines whether she adopts or not, and additional contacts are redundant, here additional contacts can increase an agent's willingness to adopt, due to later adoption events in the neighborhood. With this extension we take into account the correlation between the number of friends engaging in a behavior and the probability of adopting the behavior. The strength of a tie between two agents is here the number of common neighbors of those two agents, as in Granovetter (1978).

Let q_t^i be the *MQR* of an agent at time t. In the basic percolation model this individual threshold remains constant over time, that is to say $q_t^i = q_0^i \ \forall t$. Thus, the number of adopting neighbors does not play any role in adoption decisions. Nothing changes for an agent if she knows about the new idea from just one or many neighbors: the number of adopting neighbors does not have any weight,

[2]According to Newman and Watts (1999), the theoretical threshold value is p that solves $\mu = 1 - \frac{(1-p)^2}{4p}$.

and additional adoptions bring only redundant information. In our model we include a new factor in the expression of the individual valuation of an idea, according to which decisions are influenced by the number of adopting neighbors. Put differently, adopting neighbors can "advocate" in favor of the new idea, so as to increase the likelihood of its adoption.

The newly defined *MQR* needs to satisfy the following requirements: (a) it must be decreasing in the number of adopting neighbors, and (b) decreasing in the intensity of social reinforcement. Let $q_t^i \in [0,1]$ be the *MQR* of an agent, $a_t^i \in \mathbb{N}$ the number of adopting neighbors and $\gamma \in [0,1]$ a parameter expressing the social reinforcement intensity. For the new results to be comparable to the basic percolation benchmark, we add the following requirements: (c) for $\gamma = 0$ (no social reinforcement), $q_t^i = q_0^i \quad \forall t$, and (4) when only one neighbor adopts, $q_t^i = q_0^i$. Condition (d) guarantees that the basic percolation model is a particular case of the extended model with social reinforcement, while condition (4) ensures that the first decision to adopt (i.e. after the initial contact with the idea) is the same in both models.

We define the following *MQR* of percolation with local social pressure:

$$q_t^i = q_0^i \cdot \left(\frac{1}{a_t^i}\right)^{\gamma} \tag{3.2}$$

Proposition 1.

Function $f(a,q,\gamma) = q \cdot (\frac{1}{a})^{\gamma}$ *satisfies:*

(a) f *is decreasing in* a*, for* $\gamma > 0$

(b) f *is decreasing in* γ*, for* $a > 1$

(c) $f(a,q,0) = q$

(d) $f(1,q,\gamma) = q$

Proof.

Domain of the variables: $\gamma \in [0,1]$ *by definition;* $q \sim U[0,1] \longrightarrow q \in (0,1)$*;* f *is only evaluated after the first neighbor adopts* $\longrightarrow a \geq 1$*.*

(a) $\frac{\partial}{\partial a} f = \frac{\partial}{\partial a}(q(\frac{1}{a})^{\gamma}) = q\frac{\partial}{\partial a}a^{-\gamma} = q(-\gamma)a^{-\gamma-1} = -q\gamma(\frac{1}{a})^{\gamma+1} < 0$ *if* $\gamma > 0$

(b) $\frac{\partial}{\partial \gamma} f = \frac{\partial}{\partial \gamma}(q(\frac{1}{a})^{\gamma}) = q\frac{\partial}{\partial \gamma}a^{-\gamma} = qa^{-\gamma}ln(a)\frac{\partial}{\partial \gamma}(-\gamma) = -qa^{-\gamma}ln(a) < 0$ *if* $a > 1$

(c) $f(a,q,0) = q \cdot (\frac{1}{a})^{0} = q \cdot 1 = q$

(d) $f(1,q,\gamma) = q \cdot (\frac{1}{1})^{\gamma} = q \cdot 1 = q$ □

Neighbors always give positive information about the new idea: the more neighbors adopt, the easier it is for an agent to adopt. Social reinforcement is a positive force for adoption. With the same number of adopting neighbors, the updated value of *MQR* will be lower for higher social reinforcement intensities, so adoption will be easier. The first time that an agent makes an adoption decision occurs after the first adoption event in her neighborhood. But contrary to the basic percolation model, this decision is not definitive, since an agent can reconsider adoption at any time after she first was informed about the idea, and possibly adopt after more than one neighbor have adopted.

3.4.1 Simulation of percolation with social reinforcement

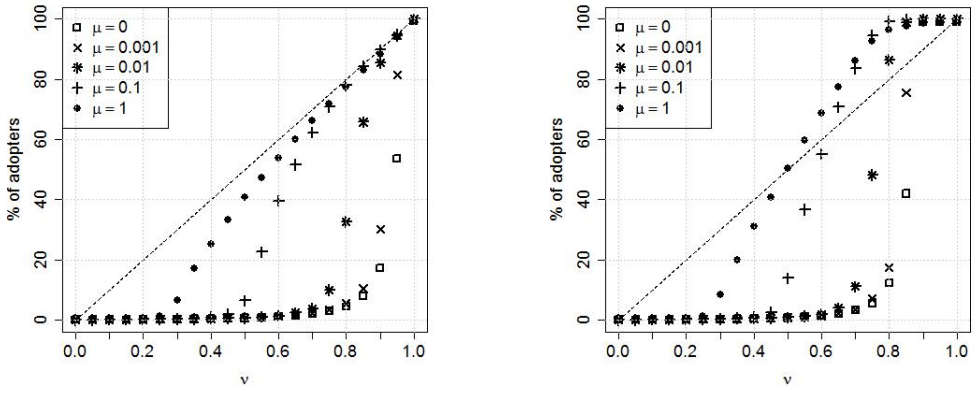
In this section we study the percolation model extended with social reinforcement by means of batch simulation experiments. The simulations are run as in the benchmark case: For the social network structure, different instances of the small world model (Watts and Strogatz, 1998) are considered, which are identified by a rewiring probability $\mu \in \{0, 0.001, 0.01, 0.1, 1\}$. We consider $N = 10{,}000$ nodes representing potential adopters, with $k = 4$ neighbors on average. We simulate the model in different settings represented by the rewiring probability μ (network structure), the initial value of the idea $v \in [0,1]$, and the social reinforcement

intensity $\gamma \in [0, 1]$. The *MQRs* of agents are random draws from a uniform distribution, $q \sim U[0, 1]$. For each setting we run $R = 50$ simulations, and look at the average value of the diffusion size across the different runs. In all simulations the diffusion process is initialized with 10 early adopters, the seeds of the simulation.

Results of the simulations are reported in Figure 3.3. Without social reinforcement ($\gamma = 0$), the social structure creates "information failures" compared to the well-mixed population with perfect information. Some willing to adopt agents never become informed of the existence of the idea because none of their neighbors have adopted it. Thus, the final diffusion size is lower than the linear demand (dashed line).

We first observe that in the diffusion regime of percolation (above the threshold represented by the sharp increase in diffusion size), the social reinforcement factor adds to the diffusion levels of the basic percolation model. This is because with social reinforcement agents get to have a subjective valuation of the idea which is above its initial value v, and can be above their *MQR* even if the initial value was below it. That is to say, some of the originally unwilling to adopt have been persuaded by their neighbors. This result is a direct implication of the *MQR* being decreasing with the number of adopting neighbors.

A second but possibly more important change is for the position of the percolation threshold. For the Poisson network ($\mu = 1$), increasing the social reinforcement intensity does very little, and thus the position of the threshold is almost unaffected across the different panels in Figure 3.3. The opposite is true for the regular one-dimensional lattice and for small world networks with $\mu = 0.001$ and $\mu = 0.01$, that see their thresholds moving substantially to lower values as γ increases. For instance, the typical small world network with $\mu = 0.01$ has a threshold close to 0.8 without social reinforcement, which goes down to about 0.7 with $\gamma = 0.4$ and to 0.6 with $\gamma = 1$.



(a) $\gamma = 0$ *(b)* $\gamma = 0.2$

(c) $\gamma = 0.4$ *(d)* $\gamma = 0.6$

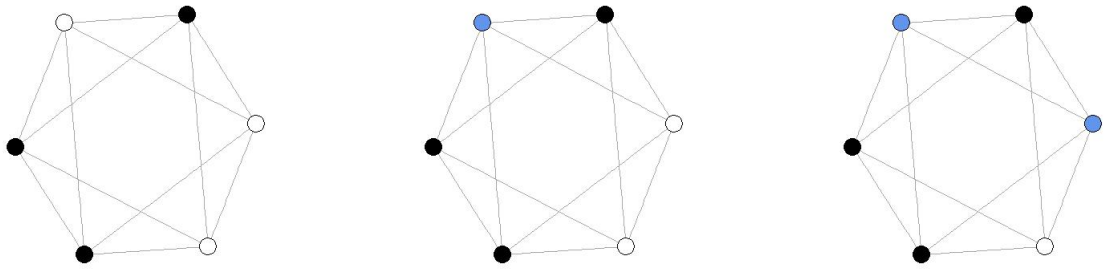
(e) $\gamma = 0.8$ *(f)* $\gamma = 1$

Figure 3.3: Diffusion size in different small world networks for different initial values $v \in [0,1]$ *of the diffusing idea (horizontal axis) in different conditions of social reinforcement intensity* $\gamma \in [0,1]$ *(different panels). Reported values are averages over 50 simulation runs. The network size is* $N = 10{,}000$ *nodes, with 10 early adopters.*

The thresholds to percolation do not just decrease, they also seem to change their nature. Without social reinforcement (Figure 3.3) the threshold from non-diffusion to diffusion regimes is a second order transition: there is a sharp but continuous change in the number of adopters. With social reinforcement (Figure 3.3), on the other hand, the threshold looks more like a first order transition: the number of adopters jumps from almost zero to almost full diffusion. As soon as there is a sufficient number of adopters, the social reinforcement forces the process to cascade to complete diffusion. This effect only happens in highly clustered networks ($\mu = 0.01$ or lower).

Without social reinforcement clustering hampers diffusion, since most links are redundant and cannot be used to reach new sources of information. With social reinforcement, though, another effect arises: shared friends may lead an agent to adoption by increasing her subjective value of an idea. This can be explained through a toy example. Assume that at a time t agent i, Alice, sees Bob, one of her neighbors, adopting the idea. Still, the initial value of the idea is below Alice's *MQR*, $v < q_r^i$. At time $t + 1$ another of her neighbors, Charlie, or agent j, adopts. This happens exactly because their common friend Bob had adopted the period before. Charlie had a lower *MQR* than Alice, which happens to be such that $v_r^j < v_{t+1} < v_r^i$. Now, with two neighbors adopting, the value of the idea for Alice becomes high enough as to be above her minimum requirement, $v_{t+2} > v_r^i$. This is how the triadic structure of their mutual friendship makes it possible for Alice to adopt at a later stage, which would have not happened in a different social structure. Figure 3.4 shows an example of this dynamic.

Moreover, Figure 3.3 shows that increasing social reinforcement intensity reduces the differences between network structures. While without social reinforcement ($\gamma = 0$) there are differences in the final size of diffusion for $v \in [0.3, 1]$ approximately, with a high social reinforcement ($\gamma = 1$) this range is reduced to $v \in [0.3, 0.5]$. This result has important implications for policies aiming at in-



(a) The white nodes represent the willing to adopt neighbors

(b) A willing to adopt agent adopts (blue node)

(c) The clustering links make social reinforcement more intense for non-willing to adopt agents

Figure 3.4: The effect of social reinforcement on clustered neighborhoods

troducing some new behavior or idea: when agents can be convinced by their friends, it is not so important to know the social network structure. In a well-mixed population, perfect information implies that every agent instantly knows about any new idea. In a network setting, this situation is represented by a fully connected network, where every agent is neighbor of every other agent. In this case, social reinforcement would lead to full diffusion even for small values of the idea.[3] Thus, it is important to know that there is some kind of network structure in the process. It is not so important, however, which structure this is as long as it is not a perfect information setting.

Finally, the simulations show evidence against the simple vs complex propagation theory. According to it, random networks perform better for simple propagations and clustered ones do better for complex propagations. Thus, random networks should do better for the case without social reinforcement ($\gamma = 0$) and worse for percolation with social reinforcement ($\gamma > 0$). Our results show that

[3]If an idea of value v is introduced, a proportion v of the N agents would immediately adopt it, that is a total of $N \cdot v$ agents. In the following step, the *MQR* of the remaining agents has been decreased by $(\frac{1}{vN})^{\gamma}$: at the end of the second step, $v(vN)^{\gamma}$ agents have adopted. The process continues so that after the s step, $v^{1+s\gamma}N^{s\gamma}$ agents have adopted. If $N > \frac{1}{v}$, then $lim_s(v^{1+s\gamma}N^{s\gamma}) = \infty$, so the process reaches full diffusion.

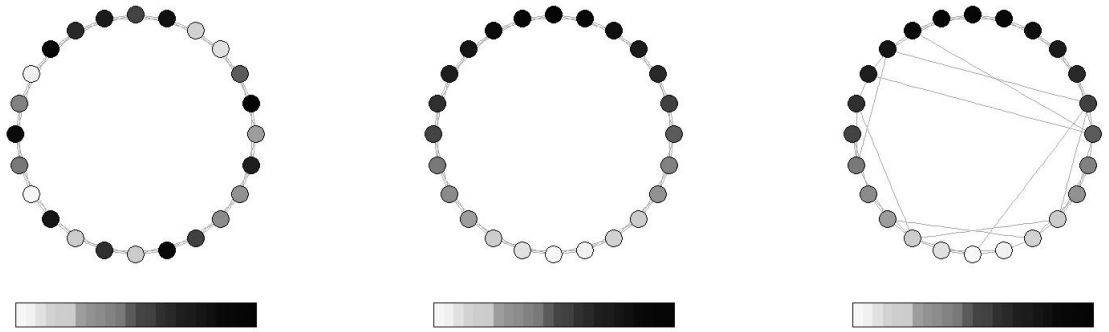
random networks outperform clustered networks both in simple and complex propagations, contrary to Centola's results (Centola, 2010, 2011; Centola et al., 2007; Centola and Macy, 2007).

3.5 Homophily

In network theory, homophily is a tendency to connections between nodes of similar characteristics. The opposite, a trend favoring connections between nodes with different characteristics, is called heterophily. A heterosexual network, for example, is a heterophilious network since partners tend to be of the opposite sex (Rocha et al., 2010). Social networks, such as friendship networks, tend to be homophilious, since they connect individuals of similar tastes or opinions (Bollen et al., 2011).

The effect of homophily and information diffusion in social networks are difficult to distinguish in empirical studies (Aral et al., 2009). The adoption of a behavior or an opinion can be both spread through the network of social contacts and due to the similar characteristics of neighboring nodes. In this section we extend the small world algorithm (Watts and Strogatz, 1998) to introduce homophily in the network.

A second problem in the study of homophily in networks is the difficulty to measure it. For categorical classes (for example, gender: an individual can be male or female), a local homophily index can be calculated as the percentage of neighbors of a node that belong to the same class of the node (Signorile and O'shea, 1964). For continuous classes, such as opinions (or, in our case, *MQRs*), an homophily index is more difficult to define. A common solution in this case is to use the continuous variable to construct a categorical classification of the nodes (in our case, the classes could be willing-to-adopt and unwilling-to-adopt nodes), as in Campbell (2013). Following this approach would lead to a loss of



(a) Original distribution of nodes in the ring lattice: the darker a node, the higher its MQR

(b) Nodes are reordered in the lattice acording to their MQR *values*

(c) After reordering, the links are rewired with probability p

Figure 3.5: Implementing homophily in the small world algorithm

information that would not allow to study the effect of social reinforcement in the homophilious network. Instead, we introduce an algorithm that allows to construct a homophilious network with nodes of similar *MQR* clustered together.

3.5.1 Modeling homophily

We introduce homophily in the model as a modification of the small world algorithm. In the original lattice, we place agents according to their minimum quality requirement and then perform the rewiring. We place the agent with the lowest *MQR*, and then the agents with the second and third lowest values by her sides, and so on. Thus, the original ring lattice is a structure of extremely high homophily. We then proceed to the rewiring, so some homophilious links are replaced with heterophilious links. The process is depicted in Figure 3.5.

The higher the number of rewired links, the lower the final homophily. In the extreme case, the random network with rewiring probability $\mu = 1$, the resulting network is equivalent with and without homophily, as the initial position of agents in the lattice is irrelevant. In the small world cases we are studying, nonetheless, this rewiring probability is not high enough to take away the ho-

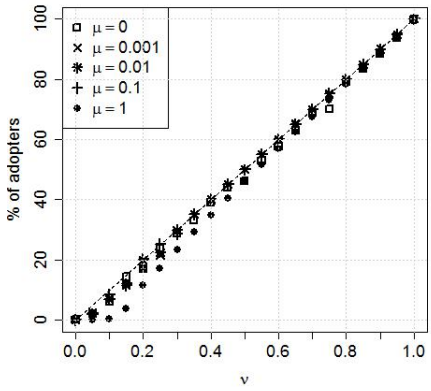
mophilious nature of the resulting network. As we consider $\mu \in \{0, 0.001, 0.01, 0.1\}$ we are rewiring at most 10% of the links: the vast majority of the links in the final network is still between agents with similar *MQR*.

3.5.2 Simulation of the homophily scenario

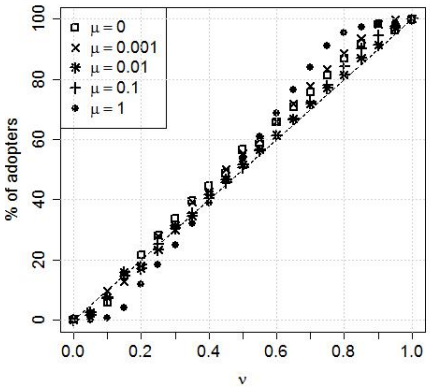
As previously, we use batch simulation experiments to study the effect of homophily on diffusion size. We consider $N = 10{,}000$ nodes representing potential adopters, with $k = 4$ neighbors on average. We simulate the model in different settings represented by the rewiring probability $\mu \in \{0, 0.001, 0.01, 0.1\}$, the idea initial value $v \in [0, 1]$, and the social pressure intensity $\gamma \in [0, 1]$. The *MQR*s of agents are random draws from a uniform distribution, $q \sim U[0, 1]$. In each setting we run $R = 20$ simulations, and look at the average value of the diffusion size. In all simulations the diffusion process is initialized with 10 early adopters.

The simulation results are reported in Figure 3.6. First, the upper-left panel shows that in the absence of social pressure, homophilious networks with different link structures present almost identical adoption sizes. In other words, the diffusion size does not differ substantially for different values of the rewiring probability. In particular, the diffusion pattern is almost the same as for a well-mixed population, that is a linear correspondence between diffusion size and value of the idea. This is equivalent to saying that the network structure of a population do not play much of a role in this context. The simulations show several main results, reported in Figure 3.6. First of all, the differences between social structures are removed: the diffusion size does not differ much between different values of μ. Without social pressure ($\gamma = 0$) they all are equivalent to a well-mixed population: we get the linear demand in all cases.

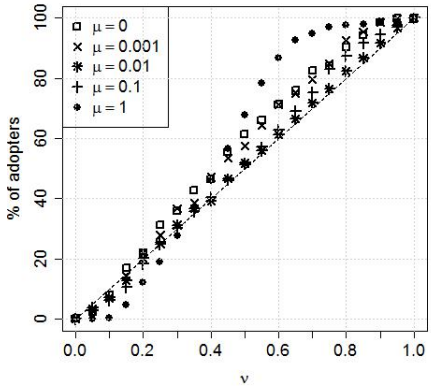
Secondly, as we introduce and increase the social pressure intensity, the diffusion sizes surpasses the linear demand, but this effect is more homogeneous than in the case without homophily (Section 3.3). In fact, the percolation thresh-



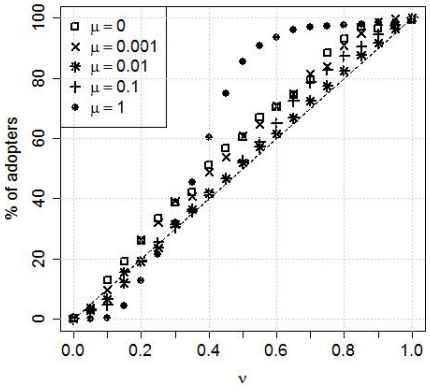
(a) $\gamma = 0$



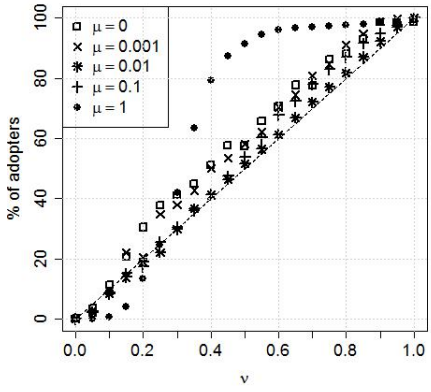
(b) $\gamma = 0.2$



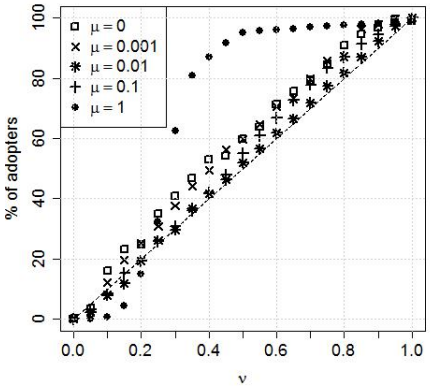
(c) $\gamma = 0.4$



(d) $\gamma = 0.6$



(e) $\gamma = 0.8$



(f) $\gamma = 1$

Figure 3.6: Diffusion size in homophilious small world networks for different initial values $v \in [0,1]$ of the diffusing idea (horizontal axis) in different conditions of social reinforcement intensity $\gamma \in [0,1]$ (different panels). Reported values are averages over 50 simulation runs. The network size is N=10,000 nodes, with 10 early adopters.

olds are lost: there is not a sharp transition any more between the diffusion and the non-diffusion regimes. Instead, there is a smooth increase in the final diffusion size as we increase the value of the idea.

Summarizing, the effect of homophily across network structures depends on the setting of the diffusion process. Without social pressure homophily always favors diffusion, as the social structure made information failures: agents that would have been willing to adopt the idea never got to know about it so never actually adopted it. The homophilious network corrects this failure, in making that willing to adopt agent are connected to one another. Once an agent adopts, all the willing to adopt become informed.

When we introduce social pressure into the picture, nonetheless, the effect of homophily is ambiguous. For low values of the idea (that is to say, in the non-diffusion regime of the non-homophilious model) the diffusion size is higher in the homophilious networks. This indicates that homophilious networks are more efficient when it comes to inform the right agents, the ones that would have adopted anyway, about the existence of idea. On the other hand, for higher values of the idea the diffusion size is lower in the homophilious networks. Those agents that were not originally willing to adopt but could be convinced via the pressure of their neighbors do not receive so much pressure to adopt. Those agents with an MQR just a bit over the value of the idea become adopters, but those agents with higher MQR do not have enough adopting neighbors to be persuaded to adopt. Thus, when social pressure is an important factor of adoption decisions, heterogeneous neighborhoods can actually boost diffusion. This is not true for the fully random network, which seems to gain a comparative advantage from social influence as compared to homophilious networks.

3.6 Non-uniform distributions: open and closed populations

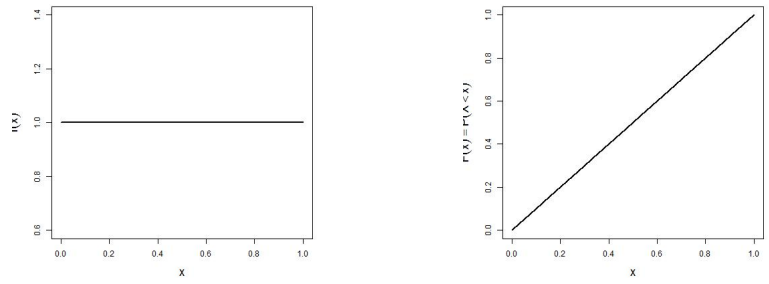
Most studies on complex propagation consider that agents are homogeneous in their resistance to contagion (Centola et al., 2007; Lu et al., 2011). The basic percolation model assumes heterogeneous but uniformly distributed characteristics of agents. However, the diffusion process is likely to depend on the specific distribution considered.

The marginal effect of an additional adopting neighbor on the *MQR* of an agent can be evaluated by differentiating (Equation 3.2) with respect to the number of adopting neighbors, as if they were a continuous variable:

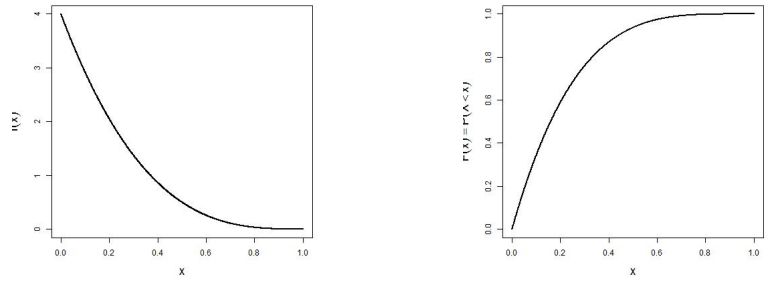
$$\Delta q_t^i = q_0^i \frac{\partial}{\partial a_t^i}(a_t^i)^{-\gamma}\Delta a_t^i = -\frac{q_0^i \gamma}{(a_t^i)^{\gamma+1}}\Delta a_t^i \tag{3.3}$$

The first adopting friends induce a large decrease in the *MQR*, while after a large number of friends have adopted the influence of an additional adopting neighbor is negligible. Moreover, the effect of one more friend adopting is larger for large values of q_0^i. In this section we study the effect of non-uniform distributions of agents' intrinsic *MQR* on percolation with social reinforcement.

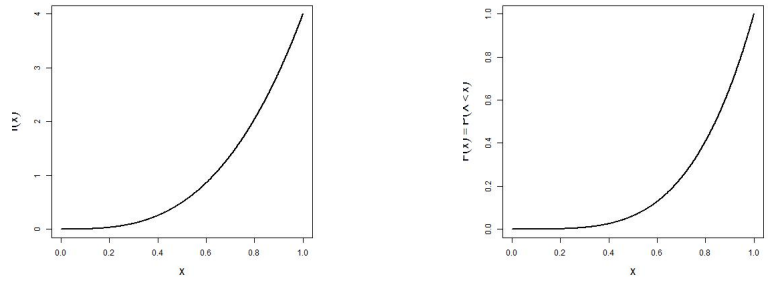
The $Beta(\alpha, \beta)$ family of distributions is particularly suited to our model, since it has support in the interval $[0, 1]$. The shape parameters α and β allow to change the mean and standard deviation of the distribution, and also to play with its symmetry. As before, the cumulative distribution indicates the potential adoption base in a fully mixed and fully informed population. Figure 7 show some examples. The uniform distribution is a particular case that one obtains with $\alpha = \beta = 1$ (panels a and b). The case of $Beta(1, 4)$ (panels c and d) describes a population where most agents have a low value of *MQR*. On the contrary, the case of $Beta(4, 1)$ (panels e and f) is for the opposite situation, where most agents



(a) Beta(1,1) = U(0,1)



(b) Beta(1,4)



(c) Beta(4,1)

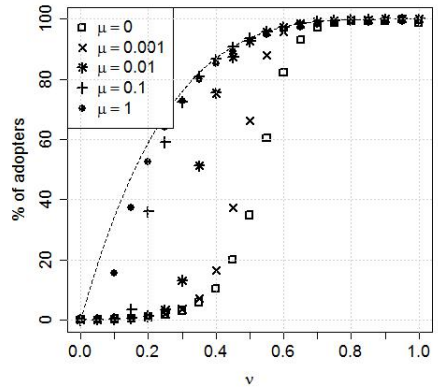
Figure 3.7: Different MQR *distributions from the* $Beta(\alpha, \beta)$ *family. Left: probability density function. Right: cumulative distribution.*

have a large *MQR*. In the context of our model, the latter case is the one of a population where most people are close-minded, or resistant to the adoption of a new idea, and only few people are enthusiastic early adopters. We may consider this one as a realistic case, where most of a population is reluctant to changes and to embrace novelties, due to reasons that span from psychology (status-quo bias), sociology (norms) to any fixed costs of investments into physical and human capital, like technological standards and infrastructures.

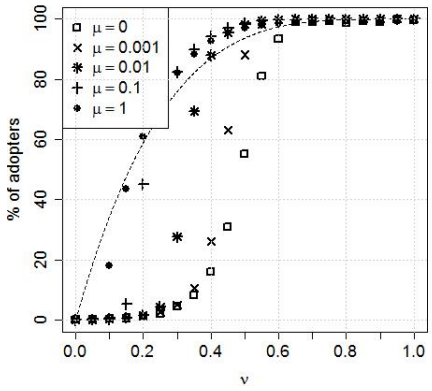
As in the previous section, we use batch simulations to compare the behavior of the diffusion process under different conditions. As before, we compare five network structures from the small world algorithm with rewiring probabilities $\mu \in \{0, 0.001, 0.01, 0.1, 1\}$, $N =$ 10,000 nodes and average $k = 4$ neighbors. We first look at an open-minded population with *MQR*s drawn from a $q \sim Beta(1, 4)$ distribution (Figure 3.8), and later to a close-minded population, with a $q \sim Beta(4, 1)$ distribution of *MQR*s (Figure 3.9). For every setting of value of the idea $v \in [0, 1]$, social pressure $\gamma \in [0, 1]$ and rewiring probability μ we study the mean diffusion size over $R =$ 50 runs with 10 seeds or initial adopters.

Similarly to the previous case, increasing the social reinforcement intensity γ increases the number of adopters, as some of the unwilling to adopt are convinced. It also lowers the percolation thresholds. In open populations most agents have a low *MQR*. Since the effect of social reinforcement is not so important for low values of *MQR*, the effect of increasing the social reinforcement intensity γ is similar to that for a population with a uniform distribution of *MQR* (Figure 3.8). For a closed population (Figure 3.9), on the other hand, an outstanding effect takes place: social reinforcement makes clustered networks (low μ) more efficient than fully random networks (large μ). This effect is due to a reinforcement of adoption provided by the redundant links of a cluster. In fully random networks the percolation threshold is less affected by social reinforcement, if at all. As the intensity of social reinforcement γ increases, the diffusion size curves of different networks become close to each other, and nearly overlap for $\gamma = 0.4$ (Figure 3.9*c*). Beyond this level, clustered networks outperform the Poisson network.

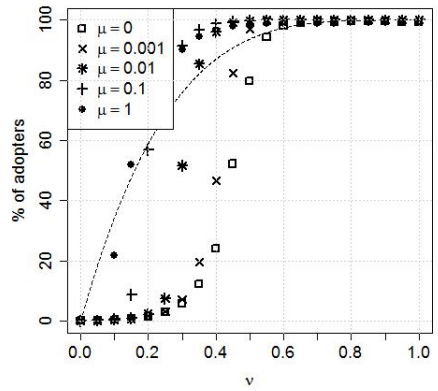
Indeed, random networks lead to higher diffusion sizes in simple propagations ($\gamma = 0$), while clustered networks perform better in cases of complex propagations (γ sufficiently large). This result sheds light on the factors that drive diffusion in social networks, and reconcile opposing empirical evidence on different diffusion



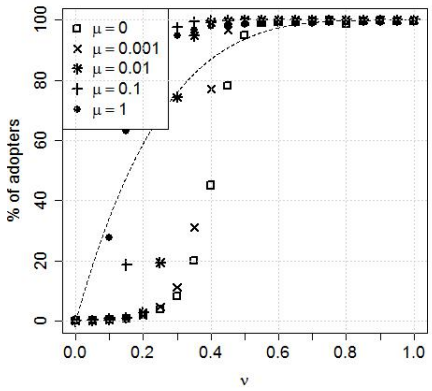
(a) $\gamma = 0$



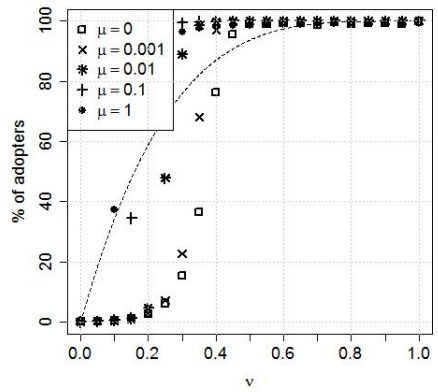
(b) $\gamma = 0.2$



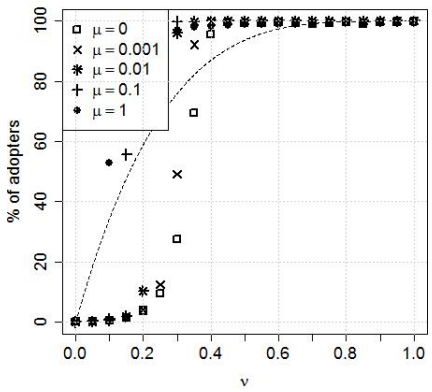
(c) $\gamma = 0.4$



(d) $\gamma = 0.6$

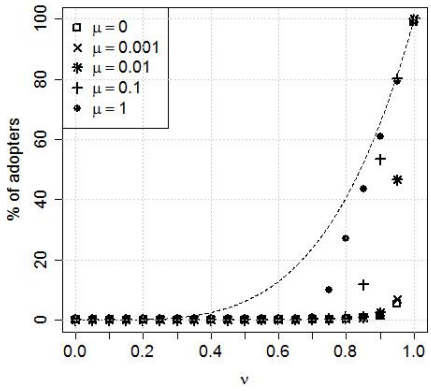


(e) $\gamma = 0.8$

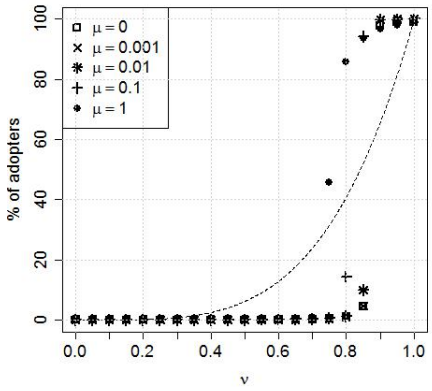


(f) $\gamma = 1$

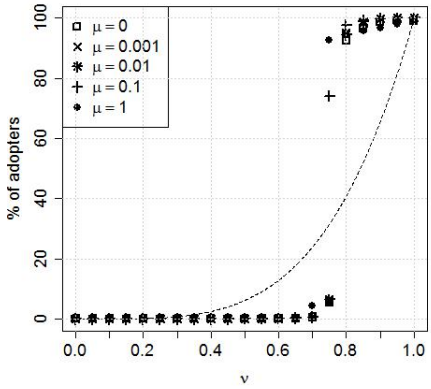
Figure 3.8: An open population: the initial MQR *values in the population follow a Beta(1,4) distribution*



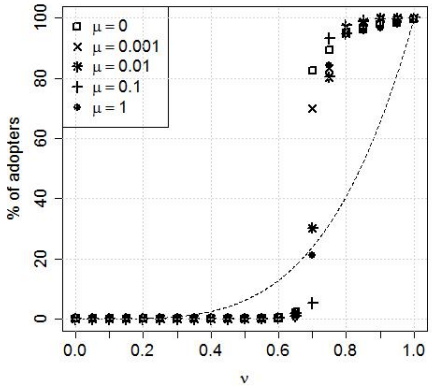
(a) $\gamma = 0$



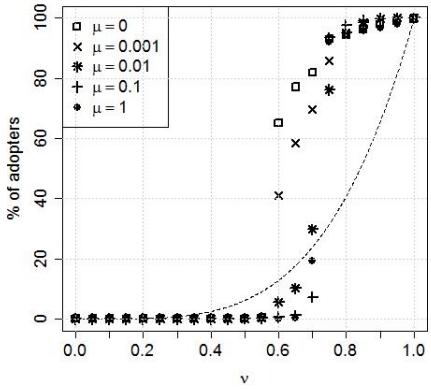
(b) $\gamma = 0.2$



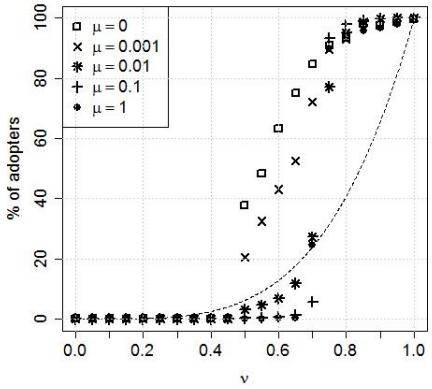
(c) $\gamma = 0.4$



(d) $\gamma = 0.6$



(e) $\gamma = 0.8$



(f) $\gamma = 1$

Figure 3.9: A closed population: the initial MQR *values in the population follow a Beta(4,1) distribution*

processes (Centola and Macy, 2007; Fogli and Veldkamp, 2012; Romero et al., 2011). Such information is highly relevant to innovation policy and development economics, in that knowing the different structure of the society under study can inform regarding the most effective policy channel to use. Broadly speaking, in a society with high dimensionality and little clustering as we have in developed countries the factor to address are individual openness levels. In developing countries instead, with highly clustered societies, it is more important to enhance the social reinforcement process in order to maximally exploit the multiplicity of close-knit links.

3.7 Conclusion

Introducing social reinforcement in a percolation model of diffusion adds to the size of diffusion. In the case of ideas that would otherwise not be diffused, social reinforcement allows for some spreading in the population. It also reduces the differences between network structures. Without social reinforcement, clustering links are redundant: if the number of ties is limited, they restrict the access to new sources of information. Nonetheless, when the opinion of neighbors can influence the adoption process, clustering links can force agents to cascade to adoption.

In simple propagations no social reinforcement is present. Thus, the size of diffusion is determined by the number of willing to adopt agents that the idea can reach. That is to say, the diffusion is determined by the dimensionality of the network, how many agents can be reached with every new step. As random networks have the highest dimensionality, they are the most efficient structures to spread an idea. In the small world algorithm, clusters come at the expenses of bridges: the more clustered the network is, the lower its dimensionality as clustering links are redundant.

For complex propagations clustering links are not redundant any more. Indeed, they provide an additional support for the social reinforcement mechanism. Once a first neighbor has adopted, the probability that a second neighbor adopts increases with the clustering coefficient of the network. In the limit case of no clustering, a random network, the probabilities of different neighbors adopting are independent. Thus, introducing social reinforcement affects the diffusion in the random network vaguely. It increases the number of adopters, as some of the unwilling to adopt are convinced, but leaves the percolation thresholds essentially unmoved. On the other hand, the interaction of social reinforcement with the structure of highly clustered networks alters both the number of adopters and the thresholds of the shift from a non-diffusion to a diffusion regime. Moreover, there appears to be a change in the nature of these thresholds, from a second order transition to a first order (discontinuous) transition. The interplay of clustered networks decreasing their thresholds while random networks remain more stable results in an homogenization of the results for the different networks. For high intensities of social reinforcement, it is important to know that there is a social network underlying the process of diffusion, but not so important to know which network it is. Nonetheless, even with this uniformization of the network structures random networks still come as the most efficient structures to enhance diffusion.

In this setting, random networks get higher shares of diffusion both for simple and complex propagations, contrary to the findings of (Centola, 2010, 2011; Centola et al., 2007). Nonetheless, changing the distribution of openness throughout the population of agents can confirm their results. Our study confirms that clustering can be favorable or harmful for diffusion, depending on the setting. Nonetheless, the determinant of which network structure is more efficient for spread is not only the nature of the process (a complex or a simple propagation), but also the characteristics of the population in which it diffuses.

CHAPTER 4

Percolation, critical fragmentation and scientific transitions

This chapter has been produced in collaboration with K Frenken and P Zeppini. The PhD candidate has been the primary researcher of the work reported in this chapter and has been the main contributor in all stages of research (idea, theory, methodology, analysis, interpretation, writing and presenting).

4.1 Introduction

Scientific transitions are generally understood as a shift in the dominant paradigm (Geels, 2010). During a transition, a new paradigm takes the place of the existing, established old paradigm and becomes dominant. Nonetheless, scientific paradigms are characterized by path dependence and the almost irreversible character of scientific development. Thus, not all alternatives to the existing paradigm will become successful, irrespective of their intrinsic benefits and possible advantages. Understanding the mechanisms that trigger a scientific regime shift can also be a policy relevant issue, whenever more or less sustainable technologies are involved.

Often societal transitions present non-linear behaviors that mimic the thresholds and tipping points of physical transitions to a large extent (Zeppini et al., 2014). This is the case of the adoption of novelties that occur as the diffusion process on a social network.[1] When considering a transition as a process of diffusion, a new theory usually must find its way in a population where the old paradigm is dominant and benefits from a large established pool of users. A critical mass of adopters is usually needed for a population to switch from an old paradigm to a new one (Arthur, 1989; Bruckner et al., 1996; Frenken and Verbart, 1998). This critical mass can come from several factors that in a way or the other reside on a coordination between adopting agents. Thus, a traditional explanation for a paradigm shift is a high concentration of adopters of the new theory in the overall population.

A fundamental aspect of scientific transitions is that a shift from an established paradigm to a new one generally requires several attempts. A number of candidate theories may fail to become dominant, before one successfully replaces the old paradigm. A usual justification is that the process of trial-and-error echoes an exploring process where new candidate paradigms learn from the failure of the previous ones. The innovative theory that succeeds in replacing the old regime is usually assumed to be better than the ones that failed, and good enough to become dominant. This simplistic view has been challenged by Brian Arthur with the concepts of increasing returns to adoption and technological lock-in: due to the positive feedback of so-called network externalities (technological infrastructure and standards, learning-by-doing, etc.) it may well be that an initially inferior technology comes out as successful one only because of favorable early adoption events (Arthur, 1989). This also applies to scientific and societal transitions in general, where behavioral sources of lock-in are to be

[1]Throughout the chapter, we talk about scientific transitions, minimum quality requirements to adoption and social reinforcement. A different interpretative framework is also possible if we designate them, respectively, technological transitions, reservation prices and increasing returns to adoption, as in Zeppini and Frenken (2015).

found in psychological features of decision making, such as the anchoring effect and the status quo bias (Kahneman, 2003).

Arthur's model as it is can not be used to describe transitions though, a strong limitation being the *ad hoc* setting where alternatives need to start competing at a given 'initial' time, a quite unrealistic feature. This very assumption prevents to unveil a number of fundamental features of a transition process. In this chapter we propose a mechanism of scientific transitions in social networks exactly focusing on the competition of different alternatives that are proposed to potential adopters at different times. Our model shows that transitions come from novelties that are 'successful' because they arrive "at the right time". The intuition of this result is as follows. Every new theory diffuses to a small amount of people, gradually shattering the ground of advocates of the old regime. Due to increasing returns to adoption, agents experience a social reinforcement in adopting the same theory of their social contacts. After several theories have failed to replace the old regime, the population is fragmented between several candidate paradigms and the old regime: the social reinforcement for the old paradigm is weakened. A new theory that comes in that moment can find the right conditions to diffuse and replace the old regime, without being intrinsically better than the previous ones. The triggering event of transitions is then a "critical fragmentation" of adopters, rather than a "critical mass", which characterizes more traditional explanations of theoretical regime shift.

We analyze our hypothesis with a model of repeated diffusion processes for a population embedded in a social network. We model diffusion in a percolation framework (Solomon et al., 2000; Zeppini and Frenken, 2015) that represents a word-of-mouth communication in a social network (Alkemade and Castaldi, 2005; Campbell, 2013). All theories have the same intrinsic value, although the population perceives them as different due to the effect of social reinforcement. Agents update their perceived value of a theory with its diffusion among their

friends compared to the number of friends that still remain in the old regime. Thus, diffusion is driven by local social influence.

We find several conditions under which our model reproduces realistic patterns of regime shift. The first condition concerns the intrinsic value of theories. If this value is too low, a new theory can never become dominant. On the other hand, theories with a very high intrinsic value will immediately replace the old regime, without a need for previous failed trials, as their intrinsic value is high enough to compensate the increasing returns to adoption of the old regime. The most realistic scenario of several attempts preceding a successful transition requires that such intrinsic value be near to the percolation threshold of the network.

A second condition is that for a transition to occur new theories need to have some advantage over older ones in order to attract adopters of failed theories. Otherwise, the population remains fragmented over different theories, none of which can become dominant. Finally, the diffusion process needs a moderate level of social reinforcement or increasing returns to adoption. If social reinforcement is too low, the number of advocates of a theory is irrelevant for its diffusion. On the other hand, too much social reinforcement can lead to a herd movement where all the adopters of a theory jump to the next arriving theory, without allowing for a fragmented population. This result depends on the intrinsic value of theories, and only holds when the intrinsic value is on the threshold of the percolation process.

In conclusion, this chapter posits a new explanation of scientific transitions under social reinforcement which is alternative with respect to traditional arguments based on trial-and-error or critical mass. We suggest that a new theory becomes dominant when arriving at the right moment, instead of being better than failed attempts. The right moment consists of a social base which is sufficiently fragmented among different competing options. Such fragmentation causes a

shift of the percolation threshold of the network towards lower values. The trigger of regime shifts is thus a "critical fragmentation" of the population, rather than a critical mass of adopters of the new theory.

The chapter is organized as follows. Section 4.2 reviews previous explanations of transitions in presence of increasing returns to adoption. Section 4.3 presents the percolation model with increasing returns to adoption and repeated entry of theories. Simulations of the model are analyzed in Section 4.4. Finally, Section 4.5 offers some concluding remarks.

4.2 Literature review

A scientific transition is a shift from a dominant paradigm to a new paradigm. Scientific paradigms benefit from social reinforcement, similarly to social norms and institutions: the benefits from adopting or complying to something increase with the number of fellow adopters, giving place to a self-reinforcing mechanism. Social reinforcement has many origins, including knowledge spillovers, economies of scale, network externalities, and learning-by-doing on the side of users (Frenken et al., 2004). Under social reinforcement, a transition requires escaping lock-in of the present theoretical system, that already counts with a great mass of adopters (Alkemade et al., 2009).

The literature on transitions has dealt with the determinants of a paradigm substitution. In a scenario with social reinforcement, the "fitness" of a theory is not only determined by its intrinsic quality, or performance, but also depends on its adoption frequency in the population. An innovative theory with a relatively high intrinsic fitness might find it hard to compete with a less fit incumbent theory that benefits from a large pool of adopters. Evolutionary models have emphasized the possibility of a theoretical lock-in into suboptimal varieties (Arthur,

1989; David, 1985). In such cases, scientific transitions become an important policy challenge.

Bruckner et al. (1996) proposed an evolutionary model of theoretical transitions where no new theory can succeed if it starts with a small number of adopters. A critical mass of adopters of the new theory is required to trigger a theoretical substitution. Since all innovative theories start with no adopters, or just a few early adopters, this model remains confined to cases where there is no established paradigm, and the critical mass is then relatively small.

Heterogeneous theories may compete on some dimension of performance or profitability. In such a case, a new theory may successfully be introduced in a population niche. This niche can protect it while it develops, until it has acquired a sufficient number of adopters to compete with the old theory (Frenken et al., 2004).

In some cases potential adopters are forward looking, and in their adoption decision they consider the potential benefit of being part of the collective adoption movement. That is, they foresee the increasing returns to adoption stemming from the coordination of fellow adopters (Lissoni, 2005). Such coordinated action provides with the required critical mass to substitute the old theory and trigger a transition.

Our chapter offers an alternative explanation of societal transitions, which builds on transitions models of physical systems. In general, a transition is a structural change of a system which involves a sizable change of some macroscopic properties following a marginal variation of a structural parameter when this reaches a critical level. Typical examples in physics are the phase transitions of materials, like the transformation of liquid water in ice. Network structures can undergo phase transitions as well, when giant connected components form out of the variation of parameters like the network connectivity. Diffusion processes on networks also present phase transitions. With percolation phenomena,

in particular, a giant percolation cluster forms when a threshold value of the 'strength' of the diffusing entity is reached.

In our model, a societal transition takes place in a social network where repeated theories that enter the market and are not successful can still steal some of the adopters of the old theory. After several failed substitutions, the old paradigm has lost part of its adopters, and the population is in a state of "critical fragmentation". No theory retains high benefits from the number of adopters, as the population is divided in their choice of theories. Under these conditions, a new theory can find the ground prepared to diffuse both among adopters of the failed theories, as among the remaining adopters of the old theory. This explanation differs from the others in that the critical fragmentation does not bring a critical mass of adopters of the new paradigm, but a critical mass of un-adopters of the dominant paradigm.

4.3 The model[2]

4.3.1 Percolation

We analyze the diffusion process of scientific theories on a population that presents a social network structure. theories are identified by their value, represented by a number $v \in [0, 1]$. Agents are heterogeneous in their incredulity, or resistance to adopt the theory.They are characterized by their minimum quality requirement (*MQR*) for adopting a new theory. The higher the *MQR* -the less open an agent is- the higher the value he requires of an theory in order to adopt it. The *MQR* of agents is a random variable, uniformly distributed $q \sim U[0, 1]$ or with any other distribution from the Beta family.

[2]This Chapter builds on previous work on percolation with social reinforcement. This Section is based on Sections 3 and 4 of Chapter 2.

The theoretical framework just presented corresponds to the so-called social percolation model (Solomon et al., 2000; Zeppini and Frenken, 2015). In this framework time is discrete, and agents adopt the new idea at any given time t if the following three conditions are met:

- the agent has not adopted before t,
- the agent is informed, which only occurs if at least one neighbor has adopted at time $t-1$,
- the value of the idea is higher than the *MQR* of the agent, that is $q_i < v$. We name those agents 'willing-to-adopt'.

In a well-mixed population, without network structure of social contacts, there is perfect information. As soon as the idea enters the society, the willing-to-adopt agents adopt, while the rest do not. Since the *MQR* is uniformly distributed as $q_i \sim U[0,1]$, a proportion $100 \cdot v\%$ of the population will adopt an idea of value $v \in [0,1]$. This case can be represented in our model with a complete network, where every agent is connected to every other agent. In a complete network, a single early adopter will inform the whole population of agents about the existence of the idea. When agents are embedded in a social network structure instead, and information travels only through social contacts, two different regimes in the *MQR* space v arise: a *diffusion* regime, where the diffusion size is about the same that one obtains in a well-mixed population, and a *no-diffusion* regime, where diffusion is almost absent. These two regions are separated by a percolation *threshold* v_c, as the result of a second-order *critical transition* (Stauffer and Aharony, 1994).

4.3.2 Social network

In a percolation setting, agents become informed of the existence of the theory through their neighbors. Thus, the structure of the social network where the agents are embedded can be determinant of the outcome of the process. Previous studies have considered percolation processes in regular networks as two dimensional lattices (Cantono and Silverberg, 2009; Hohnisch et al., 2008; Zheng et al., 2013), or in completely random networks (Campbell, 2013).

In this chapter we propose the use of the small-world algorithm (Watts and Strogatz, 1998) for the modeling of the social structure as in Cowan and Jonard (2004). This provides with a family of networks, an interpolation between regular lattices and completely random networks. The algorithm starts with a regular ring lattice and rewires every link with probability μ. This parameter allows to fine tune the randomness of the network. Figure 4.1 shows the result of simulating a percolation process in small worlds with different rewiring probabilities $\mu \in \{1, 0.1, 0.01, 0.001, 0\}$.

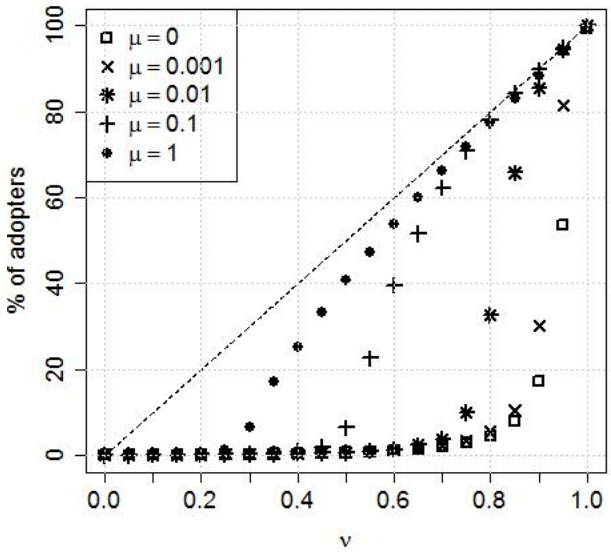


Figure 4.1: Percolation in different network structures

4.3.3 Social reinforcement

So far, only the first time the agents are informed about the theory determines whether they adopt it or not, additional contacts are redundant. Here we extend the basic percolation model with social reinforcement or local increasing returns to adoption. With social reinforcement, additional contacts will increase an agent's willingness to adopt.

In the previous chapter we included the following extension of the model to incorporate increasing returns to adoption. Let q_t^i be the *MQR* of an agent at time t. In the basic percolation model this threshold remains constant over time, with $q_t^i = q_0^i \ \forall t$. Thus, the number of adopting neighbors does not play any role in adoption decisions. We include a new factor in the expression of the value of a theory, according to which decisions are influenced by the number of adopting neighbors.

$$q_t^i = q_0^i \cdot \left(\frac{1}{a_t^i}\right)^{\gamma} \tag{4.1}$$

The updated *MQR* is defined to satisfy the following hypothesis of the model. Let $q \in [0, 1]$ be the *MQR* of an agent, $a \in \mathbb{N}$ the number of adopting neighbors and $\gamma \in [0, 1]$ a parameter expressing the increasing returns to adoption intensity. The functional form $f(q, a, \gamma)$ is chosen such that: *(a)* it is decreasing in the number of adopting neighbors, $\frac{\partial f}{\partial a} < 0$, so that the more neighbors adopt, the easier it is for an agent to adopt; *(b)* it is decreasing in the intensity of the increasing returns to adoption, $\frac{\partial f}{\partial \gamma} < 0$, so that with the same number of adopting neighbors, the updated value of *MQR* will be lower for higher γ; *(c)* with only one neighbor adopting it is equal to the initial *MQR* q_0; *(d)* in the absence of increasing returns to adoption ($\gamma = 0$) it is equal to the basic percolation model. The functional form in Equation 4.1 fulfills all four conditions.

Finally, we add a further extension to the model to account for several theories competing against the original paradigm. Let $q_t^i(j) \in [0,1]$ be the *MQR* of agent i for theory j. Once a theory has finished diffusing, a new theory enters the network to random seeds. Every neighbor that has adopted some theory alternative to the old paradigm adds to the social reinforcement of subsequent theories, according to Equation 4.2, through the effect of parameter ϕ.

$$q_t^i(j) = q_0^i(j) \cdot \left(\frac{1}{a_t^i(j) + \phi \sum_{j'=1}^{i-1} a_t^i(j')} \right)^{\gamma} \tag{4.2}$$

The different theories are all competing against the same old regime, so they are alternative theories. Parameter $\phi \in [0,1]$ measures the effect of neighbors who have abandoned the original regime: $\phi = 0$ means that the returns to adoption for theory j are only increased by adopters of the same theory j; while $\phi = 1$ means that the returns to adoption for theory j are increased by all the previous theories different from the old regime, at the same rate. If $\phi = 1$, introducing a new theory in the network is equivalent (operationally) to adding seeds to a single theory that competes against the old regime. For a single theory competing against the old regime, Equation 4.2 is equivalent to Equation 4.1.

It is important to note that in a percolation process information is spread through a word-of-mouth mechanism. Thus, individuals update their *MQR* for the adoption decision and do not receive any additional information from their neighbors after they adopt.

One of the conditions of this model is that agents can only be adopters of one theory at a time, either a new theory, either the old regime. When they face a tie, that is to say, when an agent can adopt two or more theories, they will choose their favorite theory, based on the intrinsic value of the theory, the number of neighbors adopters of that theory and the rest, and their *MQR* for that theory.

4.4 Results

In this section we study the percolation model extended with social reinforcement and repeated entry of theories by means of batch simulation experiments. Theories diffuse in a small world network of $N = 10,000$ nodes representing potential adopters, with $k = 4$ neighbors on average and rewiring probability $\mu \in \{0, 0.001, 0.01, 0.1, 1\}$. We simulate the model in different settings represented by the intrinsic value of the theories v, and the increasing returns to adoption intensity γ. The *MQR*s of agents are random draws from a uniform distribution, $q \sim U[0, 1]$.

4.4.1 A first approach to the model: the timeline of a simulation

We start the analysis of results with a depiction of how the process behaves for a single set of parameters. For this first approximation, we choose a "typical" small world network, with rewiring probability $\mu = 0.01$. Figure 4.2 shows the mean portion of agents that will adopt the new theory depending on the intrinsic value of the theory v and the increasing returns to adoption intensity γ, over 20 Monte Carlo runs. The colors show the standard deviation over the different runs. The zones with higher standard deviation show the thresholds between the diffusion and the non-diffusion regimes. We will use a combination of parameter values in these thresholds, where the result of the process is most unpredictable a priori.

Figure 4.3 shows the result of simulating the process with an intrinsic value of $v = 0.6$ and increasing returns to adoption intensity $\gamma = 0.6$. This combination of parameters is in the threshold of the diffusion and the non-diffusion regimes shown in Figure 4.2. Every line represents the percentage of adopters of a different theory that tries to replace the old regime. For the first half of the simulation period, many theories enter the population and fail to diffuse. They all behave

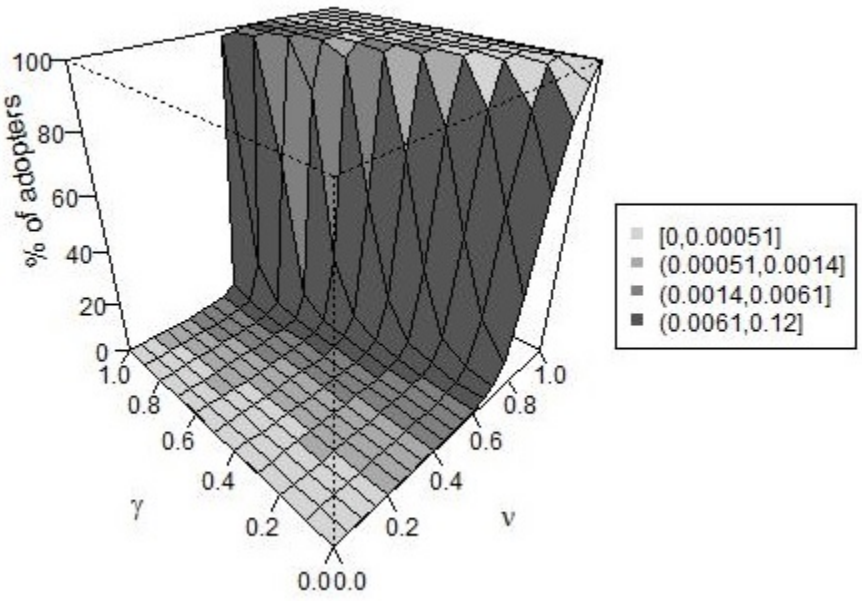


Figure 4.2: Diffusion of a single theory, depending on the intrinsic value v and the increasing returns to adoption intensity γ. The rewiring parameter is $\mu = 0.01$.

similarly, getting to a similar portion of the population. At around period $t = 500$, a new theory enters the population that fares significantly better than the rest (the pink line). This theory marks the beginning of a different pattern of diffusion. From that point on, every new theory diffuses to a higher portion of the population than the last, collecting adopters both from earlier theories as from the old regime. Finally, the last theory becomes dominant in the population and only competes against the old regime.

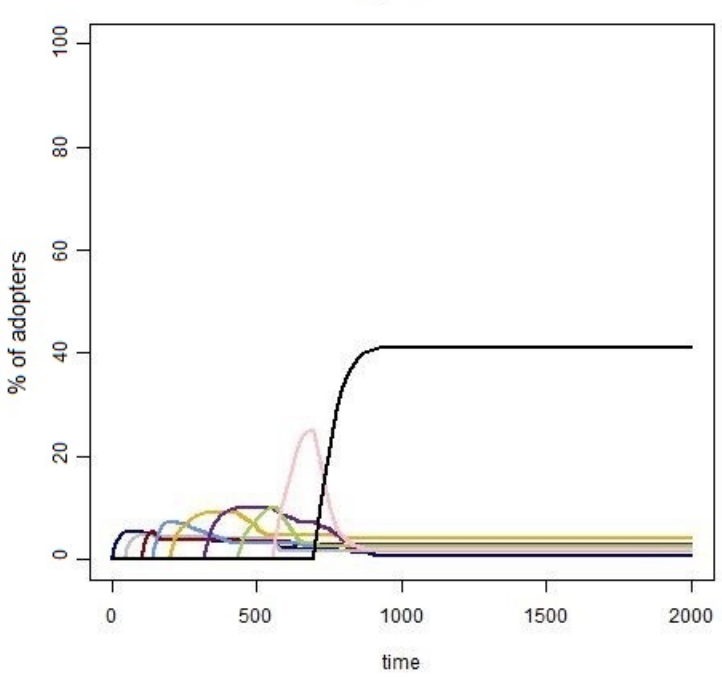
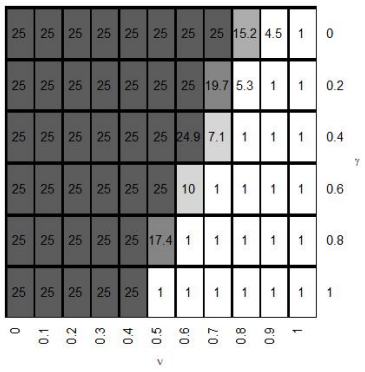


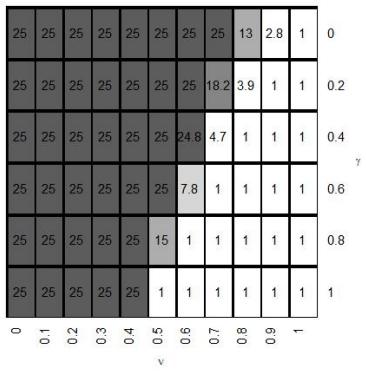
Figure 4.3: Time series of the number of adopters for different scientific theories ($\mu = 0.01$, $v = 0.6$, $\gamma = 0.6$): critical fragmentation triggers diffusion.

4.4.2 Simulations of the model

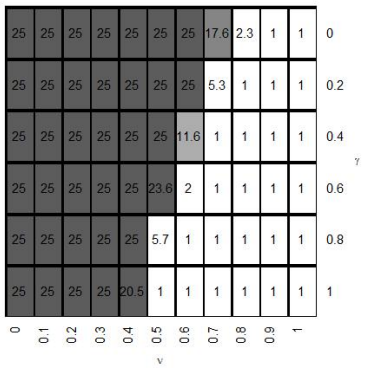
Figures 4.4, 4.5 and 4.6 show the result of running Monte Carlo simulations of the model. We consider that a transition takes place when at least half the population has adopted the a same theory alternative to the old paradigm. A scenario is defined as the simulation over a combination of rewiring probability μ, value of the theory v and increasing returns to adoption intensity γ. Figures 4.4, 4.5 and 4.6 compare the mean number of theories that had been introduced before the transition takes place for every scenario over $R = 10$ runs, with a maximum of 25 theories and 2,000 time steps. For every rewiring parameter μ, the number in the cell indicates the number of ideas before the transition takes place, for a combination of value v (vertical axis) and increasing returns to adoption intensity γ (horizontal axis). Darker cells indicate that more theories were needed.



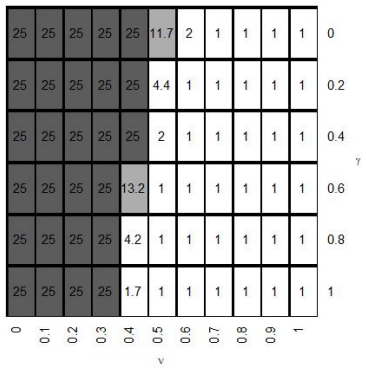
(a) $\mu = 0$



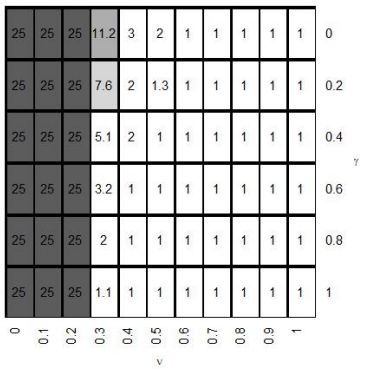
(b) $\mu = 0.001$



(c) $\mu = 0.01$

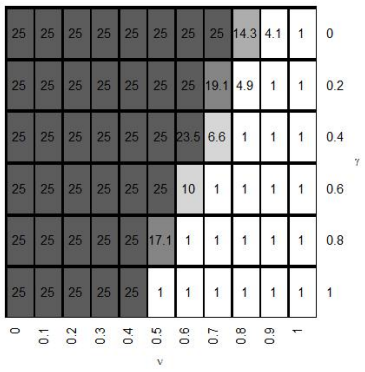


(d) $\mu = 0.1$

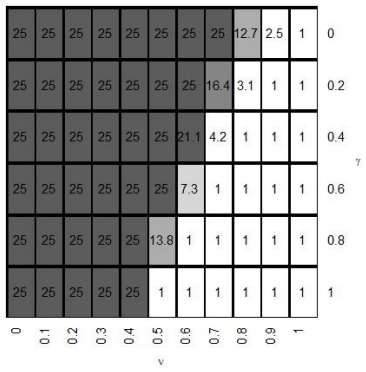


(e) $\mu = 1$

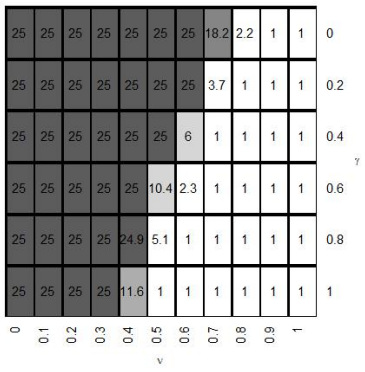
Figure 4.4: Transitions for different network structures, $\phi = 0.2$*. For every rewiring parameter* μ*, the number in the cell indicates the number of ideas before the transition takes place, for a combination of value* v *(vertical axis) and increasing returns to adoption intensity* γ *(horizontal axis). Dark cells indicate no transition.*



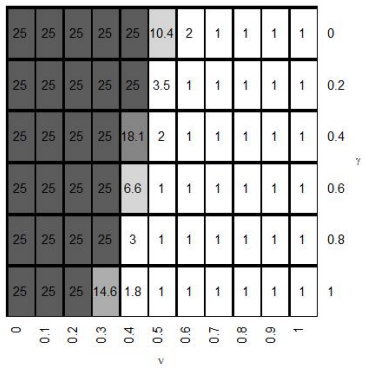
(a) $\mu = 0$



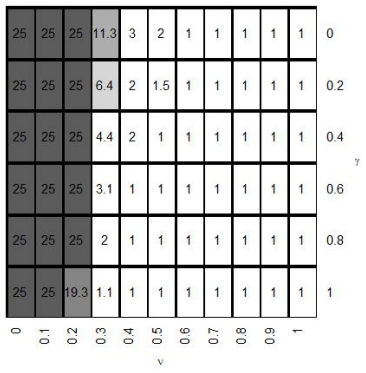
(b) $\mu = 0.001$



(c) $\mu = 0.01$

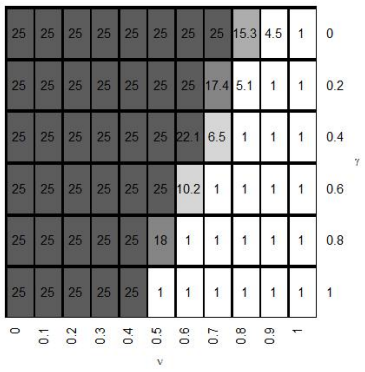


(d) $\mu = 0.1$

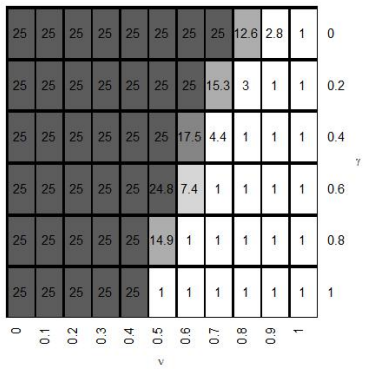


(e) $\mu = 1$

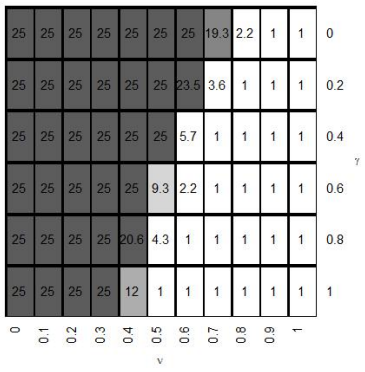
Figure 4.5: Transitions for different network structures, $\phi = 0.5$*. For every rewiring parameter* μ*, the number in the cell indicates the number of ideas before the transition takes place, for a combination of value* v *(vertical axis) and increasing returns to adoption intensity* γ *(horizontal axis). Dark cells indicate no transition.*



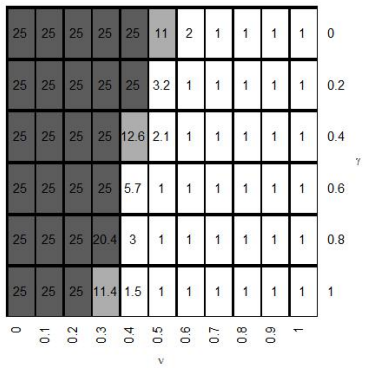
(a) $\mu = 0$



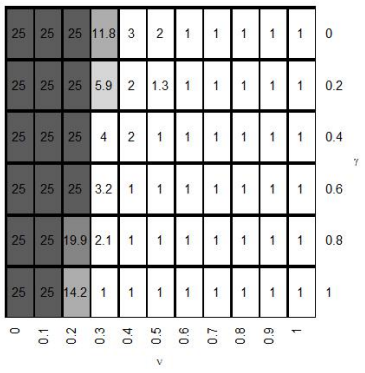
(b) $\mu = 0.001$



(c) $\mu = 0.01$



(d) $\mu = 0.1$



(e) $\mu = 1$

Figure 4.6: Transitions for different network structures, $\phi = 0.8$*. For every rewiring parameter* μ*, the number in the cell indicates the number of ideas before the transition takes place, for a combination of value* v *(vertical axis) and increasing returns to adoption intensity* γ *(horizontal axis). Dark cells indicate no transition.*

Black cells indicate that no transition took place, while white cells mean that a single theory was enough to replace the old paradigm. Grey cells are the interesting ones for our analysis: they signal that after several theories failed to become dominant, one theory became dominant in the network. For example, in Figure 4.4*c*, 11.6 theories are needed in average for a transition to take place for value $v = 0.6$ and social reinforcement intensity $\gamma = 0.4$. That is to say, the eleventh theory (on average) becomes dominant in the network.

Increasing the effect of additional theories on diffusion, parameter ϕ, does not change the results qualitatively. Indeed, Figures 4.4 ($\phi = 0.2$), 4.5 ($\phi = 0.5$) and 4.6 ($\phi = 0.8$) display a qualitatively similar behavior of the interaction of the structure of the network with the value of the theories and the social reinforcement intensity.

Moreover, as the rewiring parameter of the small world algorithm increases, the social reinforcement intensity becomes less relevant: the dark cells are situated in the top left triangle for low values of μ, while they are at the left side for high values of μ. That is to say, for the regular ring lattice (Figures 4.4*a*, 4.5*a* and 4.6*a*) the social reinforcement intensity γ interacts with the value of the theories v to determine the number of theories needed for a transition. For the random network (Figures 4.4*e*, 4.5*e* and 4.6*e*) the number of theories needed for a transition remain fairly constant for the different values of γ.

4.5 Conclusion

During a theoretical transition, a new theory takes the place of an already existing dominant paradigm. A transition can be analyzed from a diffusion angle, considering that a new theory has to diffuse in a population where the old paradigm is dominant and benefits from increasing returns to adoption. Traditional explanations of transitions under increasing returns to adoption usually require a critical

mass of adopters of the new theory to trigger a regime shift. Such a critical mass can form in a market niche, or from coordination between agents.

This chapter suggests an alternative explanation of transitions under increasing returns to adoption. Instead of a critical mass of adopters, that is to say, a critical level of coordination against the old regime in the population, we consider the effect of a critical fragmentation, that is to say, a critical level of discoordination against the old regime. Every new theory that fails to replace the dominant paradigm, gradually shatters the pool of adopters of the old regime. After several tries, the population is fragmented between several candidate paradigms and the old regime: the social reinforcement for the old paradigm is weakened. The conditions in such a case are propitiate for a new theory to gather the adopters of the previously failed theories, as well as those from the old regime, and become the new dominant paradigm.

This hypothesis is analyzed by means of a simulation model repeated diffusion processes for a population embedded in a social network. Diffusion is modeled with the percolation framework (Solomon et al., 2000), extended with increasing returns to adoption. All theories have the same intrinsic value, although agents update their perceived value of a theory depending on the number of fellow adopters in their local environment, compared to those that still remain in the old regime.

The simulations show that critical fragmentation can shatter the increasing returns to adopting the old regime and build those favorable to the new theories, under certain conditions. First of all, the intrinsic value of the theories and the increasing returns to adoption intensity must be in the threshold between the diffusion and the non-diffusion regimes. That is to say, in the non-diffusion regime, none of the theories that enter the population will diffuse, while in the diffusion regime, every new theory will replace the previous one (including the

old regime) and become dominant. In the threshold, theories will build on each other to replace the old regime.

Finally, the intensity of increasing returns to adoption has to be moderate. If it is too low, the number of fellow adopters does not affect the process. On the other hand, if it is too high, the population cannot stay fragmented. Every new theory replaces the previous one and diffuses to a bigger portion of the population. With a moderate intensity, increasing returns to adopting a new theory, and unadopting the old regime, build slowly with theories that fail to diffuse until they are high enough for a new theory to absorb the adopters of the previous ones and diffuse over those of the old regime.

In conclusion, this chapter postulates a new explanation of theoretical transitions under increasing returns to adoption. We suggest that a new theory becomes dominant by arriving at the right moment. The right moment, in this case, is after several failed theories have fragmented the social base of the old paradigm. Such fragmented population of adopters of different competing theories causes a shift of the percolation threshold of the network towards lower values. The trigger of a regime shift is thus a "critical fragmentation" of the population, rather than a critical mass of adopters of the new theory.

CHAPTER 5

Sleeping beauties in technology: delayed recognition of breakthrough innovations

This chapter has been produced in collaboration with K Frenken and JM Azagra-Caro. The PhD candidate has been the primary researcher of the work reported in this chapter and has been the main contributor in all stages of research (idea, theory, methodology, analysis, interpretation, writing and presenting).

5.1 Introduction

Ever since the first studies on innovation as the motor of economic development, scholars have distinguished two sizes of the inventive step: incremental and radical. Incremental innovations are based on small improvements along an established technological path, while radical innovations involve a deeper change, and usually imply a change of the established technological paradigm. It is generally accepted that these radical innovations, also indicated as breakthroughs, are the key for economic growth.

Breakthroughs are disruptive innovations that involve major changes in the dominant technological trajectory. The established technological community, thus, might not be able to recognize at first their implications and follow for a

while the conventional, well-known technological path, even after it has been rendered obsolete by a breakthrough development. Possible explanations for this tendency include path dependence of the technological process (David, 1985), the lack of technologically complementary technologies (Valentin and Jensen, 2002), or a a weak social position of the inventor (Singh and Fleming, 2010).

Understanding why some breakthrough innovations experience delayed recognition would help unveiling possible failures in the process of technological diffusion. A new advance at the frontier of technological progress may be ahead of its time and remain latent until complementary knowledge that builds on it has been developed. An alternative hypothesis would be that the social network of inventors is determinant for the diffusion of inventions and isolated actors with a weak social position lack the means to make their inventions noticed. This second explanation would reveal a shortcoming of the technology system, since important developments are ignored, delaying further technological development. In such a case, there may be scope for policy action to correct this flaw in the diffusion of technologies.

This study explores the case of sleeping beauties (SBs), breakthrough inventions that experienced delayed recognition, by means of patent data. Patents are legal documents that grant exclusive rights of exploitation of an invention for a limited period of time, in exchange for public disclosure. They include one or more claims, which define the protection conferred by the patent, and usually contain references to earlier patents or scientific documents. References in a patent signal the state of the art on which the patent is based, and they can limit the property rights established by its claims. A patent that is cited in many others, thus, certainly includes some technology central to further developments. Thus, patent citations can be used to study patented breakthrough innovations, identifying them as highly cited patents (Castaldi et al., 2015; Singh and Fleming, 2010). In this chapter we will add to this strand of literature by analyzing the

determinants of a delay in the recognition of breakthroughs by studying highly cited patents that did not receive any citations for a very long period after their priority date.

The structure of the chapter is as follows. Section 5.2 summarizes the previous work on breakthrough innovations and delayed recognition. Section 5.3 introduces the methodology of the study, and Section 5.4 presents the main characteristics of SBs. Finally, Section 5.5 provides some conclusive remarks.

5.2 Literature

5.2.1 Breakthrough innovations

Literature on innovation defines two sizes of inventive steps: incremental and radical. While the difference between them can be difficult to measure in practice, it is widely accepted that radical innovations are the core of technological progress and wealth creation (Ahuja and Lampert, 2001; Schumpeter, 1934). Many studies try to identify the determinants of breakthroughs in order to understand the underlying mechanisms of technological progress.

Breakthrough inventions have traditionally been linked to lone inventors, since individuals prefer to create new systems rather than to improve the systems of others (Hughes, 2004; Schumpeter, 1934). Several problems arise in creative teams, such as idea blocking, communication problems, and personal tensions (Mullen et al., 1991; Paulus and Nijstad, 2003). These problems are also linked to work in big firms, often tied by bureaucracy and the constraint to look for immediate benefits (Ahuja and Lampert, 2001). Nonetheless, the best environment for breakthroughs is unclear. Jewkes et al. (1969) signalled that breakthrough inventions originated in large corporations tend to come from the work of a single outstanding individual. Nonetheless, due to the increasing costs of R&D, large corporations provide the most fruitful source of inventions (Jewkes et al., 1969).

On the other hand, other studies on teamwork have also found evidence that collaborative research is less likely to produce very unsuccessful outcomes, and at the same time more likely to produce very successful ones (Singh and Fleming, 2010).

Another source of variability lays in the scientific origin of inventions. Several studies point out that basic science is a requisite for breakthrough invention in many technological areas (Malva et al., 2015; Melese et al., 2009; Valentin and Jensen, 2002). The reason is usually that breakthroughs are likely inventions that reside outside the boundaries of the current paradigm, so unveiling them requires a trajectory shift (Malva et al., 2015). Thus, basic scientific capabilities are more likely to generate the unexpected outcomes that lead to a breakthrough invention (Sobrero and Roberts, 2001).

5.2.2 Delayed recognition and sleeping beauties in science

Sleeping beauties in science are articles that go unnoticed for a long time and then, almost suddenly, attract a lot of attention. They were first named by van Raan (2004), although the phenomenon of delayed recognition had been known to science for a long time (Barber, 1961; Cole, 1970; Stent, 1972). There can be several reasons for a paper to become a sleeping beauty (Li et al., 2014): for instance, publishing in the wrong journals (targeting an audience that is not interested), the reputation of the author, or presenting a groundbreaking theory. This last is the case of researchers that were ahead of their time.

In science, a researcher can be ahead of her time if she publishes a theory inconsistent to the established theory (the case of Mendel) or if she researches in an early field (the case of string theory presented by van Raan (2004)). This effect is reinforced when the research cannot be continued due to a delay in technological discovery. Indeed, Lachance and Lariviere (2014) found that sleeping beauties behave differently in technical sciences (biology, physics, etc.) than in

arts and social sciences. In arts, humanities and social sciences, sleepers had the strongest start (received many more citations in the early period after awakening) and the weakest finish, while the reference group of non-sleepers had the weakest start, but the strongest finish.

Sleeping beauties are a particular case of delayed recognition. A paper suffers from delayed recognition if its peak of citations starts later than usual. Sleeping beauties are not only sleepers (the equivalent to delayed recognition in our metaphor), they also need to be beauties. That is to say, they have to be highly cited in their awakening period.

5.3 Methods

5.3.1 Patent data

Patents have been extensively used to measure innovation. Albeit all their flaws, they are very detailed information, and patent applications have two key requirements to be granted, novelty and applicability, that make them extremely useful as innovation indicators. Moreover, patent citations have been linked to patent value in numerous empirical studies (i.e. Albert et al., 1991; Harhoff et al., 1999; Sampat and Ziedonis, 2005), since Trajtenberg (1990) pointed them out in their seminal paper. Moreover, citations to patents have also been used to identify breakthrough inventions.

A breakthrough invention is usually defined as one that provides a framework for plenty of subsequent inventions, contributing disproportionally to further technological development (Fleming, 2001). References in a patent have legal binding, and they summarize the state of the art on which the patent is based (Hall et al., 2005; Harhoff et al., 2003). A patent cited in the references list of many other patents is relevant to many posterior inventions: it is likely to be a breakthrough invention.

The data on patents is extracted from the Patstat (October 2012) database. We consider patent families, together with the priority date (first application date in the family), as units of study. Considering patent families allows to group those patents that have been applied for in several patent offices, reducing the number of duplicates in the citations lists and produces more robust measures of impact (Bakker et al., 2016). As we are dealing with highly cited patents, we only consider those patents families that have received at least one citation over their lifetime.

The initial database is formed of 14,835,792 patent families, with priority years ranging from 1782 to 2011. In order to limit the variation in the data due to changes in the structure of society, we consider only families with recent priority dates. We use as a cutting point the year of creation of the World Intellectual Property Office (WIPO), 1967. Nonetheless, other relevant dates could have been chosen.[1] As the number of patents published per year increases over time and earlier years are poorly covered, the final number of patents is 11,913,774 families, 80% of the database, despite using only 45 years out of several centuries of data.

All citations from the patents in one family to the patents in another family are only counted once, as one citation from the citing family to the cited family. Due to legal issues in the lifetime of patents, sometimes a patent family can cite a family with a later priority date. These cases are not very common (they only account for 579,228 out of 75,040,261 total citations), so we dismiss them for practical reasons.

[1]Results remain qualitatively identical for different starting dates, including 1782 (the first year in the database), 1972 (the first quartile of the distribution of priority years), and 1978 (the year of the first PCT).

5.3.2 Defining sleeping beauties

The distribution of the number of citations that patents receive is depicted in Table 5.1. The distribution of citations is, as expected, close to exponential: out of the patents that have received at least one citation, more than a third have received exactly one citation, and more than half of them have received two or less citations. Following the tradition in breakthrough research (Ahuja and Lampert, 2001; Castaldi et al., 2015; Singh and Fleming, 2010), we define highly cited patents as those in a top quantile of the distribution. For this analysis, we chose the top 10% of the citation distribution, because it is quite a robust number across different initial years of the database.[2] That is to say, we consider that a patent is highly cited if it has received at least 13 citations over its lifetime.

Sample fraction	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Nr of citations	1	1	1	1	2	2	3	4	7	12	2,363

Table 5.1: Quantiles of the citations distribution

We define that a patent is sleeping for a year if it does not receive any citations for that year. Thus, a patent is sleeping until it receives a citation. We could relax this definition by including years in which it received one (or few) citations, as van Raan (2004) did with scientific papers. Nonetheless, given that most patents receive very few citations, allowing for citations during the sleeping period would transform most of the patents in sleepers. The distribution of sleep lengths (the number of years before the first citation) is shown in Table 5.2. Some patents receive their first citation during the same year of their application. For half of the patents, the first citation arrived 3 or less years after their priority date.

[2]The 90% quantile of the citation distribution is 12 citations, regardless whether the initial years of the database is 1782, 1972, 1967 or 1978.

As before, we define a sleeper as a patent family in the top 10% of the sleep length distribution. That is to say, a patent is a sleeper if it went uncited for at least 13 years after its priority date.[3]

Sample fraction	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Years of sleep	0	1	1	2	3	3	4	6	8	13	45

Table 5.2: Quantiles of the length of sleep distribution

A sleeping beauty is defined as a patent family that is both a sleeper (is not cited for at least 13 years after its priority date) and a highly cited patent (receives at least 13 citations). According to this definition, the database contains 3,196 sleeping beauties. Their priority years range from 1967 to 1996, almost the whole range of 14-year-old patent families in our database.[4] Their sleep periods range from 14 to 38 years, and their total number of citations range from 13 to 164. Figure 5.1 shows the distribution of SB families across sleep lengths and number of citations.

Figure 5.1 shows that SBs are not uniformly distributed. Those in the right side of the table are more cited ("more beautiful"), while those in the lower rows have experienced longer delay in their recognition (was “more sleepy”). Moreover, there are many more SBs in the left top corner than in the rest of the table. Ke et al. (2015) have suggested some kind of a “SB index”, a measure of how beautiful and sleepy a document is. According to such an index, most of the patents in our study would barely be SBs, while some of them (like the patent that received 55 citations after 36 years sleeping, or the one with 138 citations that slept for 24 years) would score very high. As one changes the definition

[3]The 90% quantile of the distribution of sleep length varies from 20 years of sleep if the database starts in 1782, to 13 (1967), 12 (1972) and 11 (1978).

[4]Since the database includes 2011, sleeping beauties could have had priority year 1997 in theory. Nonetheless, there was no patent from 1997 that received no citations until 2011 and more than 12 citations in 2011.

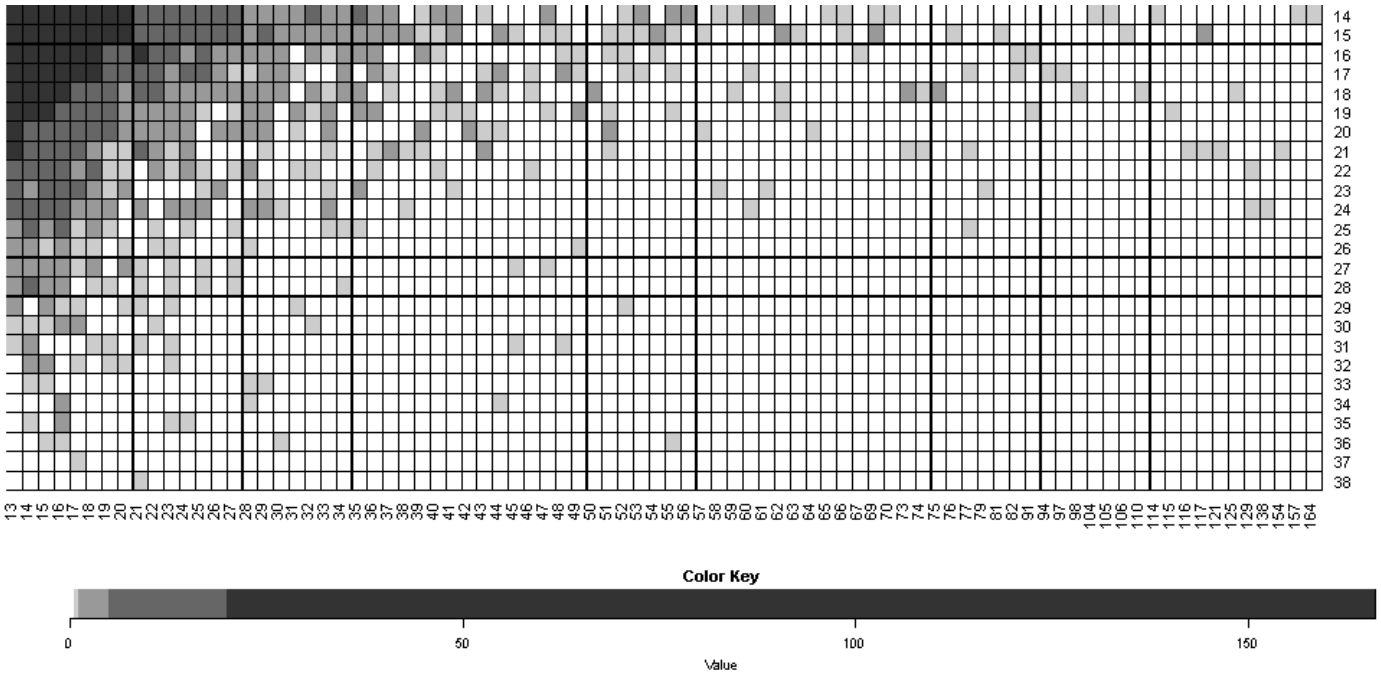


Figure 5.1: Distribution of sleeping beauties over sleep length (y axis) and number of citations (x axis). Darker shades indicate higher frequencies.

of SBs to higher degrees of this index, less documents fulfill the SB condition, which accounts for the darker top-left corner of Figure 5.1. In fact, Ke et al. (2015) found that this distribution of quantity of SBs in science depending on the index follows a power-law distribution. For a further study we could further group SBs in several levels, according to the total number of citations and the length of the sleep.

Some studies (van Raan, 2004, for instance) have defined a sleeping beauty not depending on the lifetime citations but depending on the number of citations during a fixed period after awakening. That would require the definition of an additional parameter, the length of the awakening period. Theoretically, this would allow a more accurate comparison between the older and newer patents. Nonetheless, 12 citations comes as a very consistent 90% quantile of the number of citations overall, independently of the starting year of our database. Thus, we feel pretty confident that those patents with more than 12 citations are the highly

cited patents. Moreover, all SBs are grouped together, without consideration for their intensity: we consider that a sleeping beauty is a patent that slept for 14 years and received 13 citations as well as a patent that slept for 24 years and received 154 citations. Thus, defining different awakening periods will not change their identification in our study.

In the following section, we compare the population of SBs with a control group of non-sleeping highly cited patents (HCPs).[5] Thus, we are not trying to explain what makes a patent become highly cited, but what makes a highly cited patent sleep before recognition. The question is not whether SBs differ from the whole population of patents, but rather if they differ from the population of other (non-sleeping) highly cited patents. Thus, we compare them with a control group of highly cited patents that are not sleepers. Furthermore, we consider only patents in the same period as the SBs, from 1967 to 1996. The control group is formed by 824,920 HCPs.

5.4 Results

5.4.1 Descriptive analysis

This section compares the population of SBs against the population of HCPs, across several characteristics: technological classes, number of references, and number, country, and experience of of the authors. These univariate comparisons, are a first, exploratory approach to the empirical analysis of the determinants of SBs. Results of this exploratory analysis have to be taken with caution since they are unidimensional comparisons. In the next section we will introduce all dimensions in a multivariate analysis. Results shown in this section are all for full counts, but they do not change qualitatively if we consider fractional counts.

[5]Since SBs and the control group form the whole population of patents with more than 12 citations, an alternative name for the control group could be "beauties that did not sleep", or non-sleeping beauties (NSBs).

The first dimension through which we compare SBs and HCPs are technological classes, measured through IPC codes. IPC codes in patents are assigned by the examiners, based on potential applicability. They are grouped in eight main classes, coded from A to H: human necessities (A), performing operations and transporting (B), chemistry and metallurgy (C), textiles and paper (D), fixed constructions (E), mechanical engineering, lighting, heating, weapons and blasting (F), physics (G) and electricity (H). Here, we expect that patents from technology classes that are more science-based are less likely to be SBs, because knowledge in these fields is more codified, allowing it to diffuse more freely and rapidly (Foray, 2004).

Figure 5.2 shows the distribution of SBs (black bars) and HCP (grey bars) across the eight technological classes, in percentages of the full count (every patent can be assigned to one or more IPC, so that patents with many technological classes are counted once in every class), although the results are not qualitatively different for the fractional counting case. To compare both distributions, we run a χ^2 goodness-of-fit test. It confirms ($X^2 = 488.6459, df = 7, p-value < 0.001$) that both distributions are statistically different. SBs are more probable than HCP in most technological classes, except in classes G and H, that is to say physics and electricity. Physics (G) and Electricity (H), together with chemistry and metallurgy (C), are the most science-based technological classes (Azagra-Caro et al., 2006). This comparison suggests that breakthroughs in physics and electricity are more likely to be recognized as such earlier in their lifetime.

A second dimension of comparison is the number of references cited in the patents. Since the list of references in a patent can limit the protection of its claims, one would expect inventors and applicants to cite as little prior patents as possible. In order to correct this incentive not to disclose information, in some patent offices (like the USPTO) an opposition to a patent will be more likely to be successful in court if the application of the opposed patent did not include a cita-

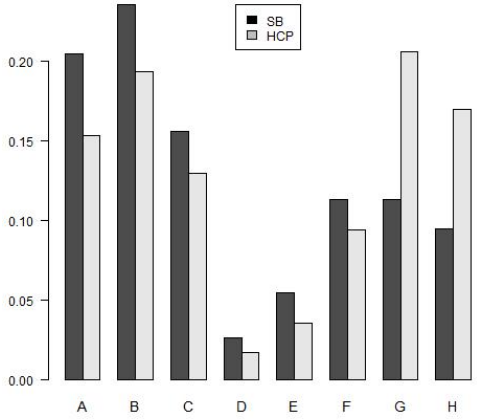


Figure 5.2: Technological classes

tion to the opposing patent. Thus, the incentives are balanced for an applicant to disclose the list of prior knowledge of the patent. Figure 5.3 shows the cumulative distribution of the number of references per patent for the SBs (black line) and the HCP (grey line) groups. A Mann-Whitney U test ($W = 451183786, p - value < 0.001$) shows that both distributions are significantly different. SBs tend to have less citations than HCP: half the SBs have 4 references or less, while half the HCP have less than $7 - 8$ references, almost the double. This could indicate that SBs are breakthrough inventions outside the existing technological paradigm: less related to existing technologies, and therefore, with less references.

Another dimension in which we expect differences between SBs and HCP is the experience of the authors (inventors or applicants). Experienced authors that have filed many applications (or made many inventions) are more known in the technological community. Hence, one can expect that their inventions are noticed earlier. A group of authors will likely profit of the experience of the most experiences one. Thus, we consider the experience of the authors of a patent as the highest in the team. That is to say, the experience of the inventors of a patents is equal to the maximum number of patents invented by any of its

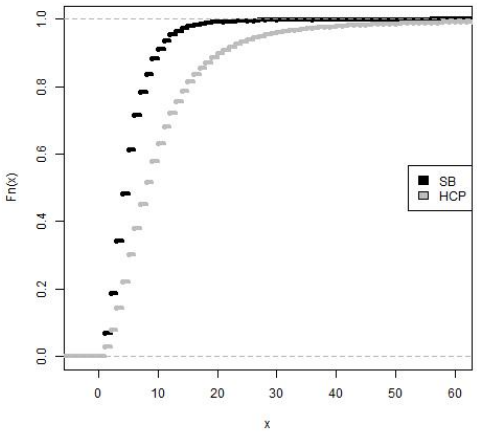
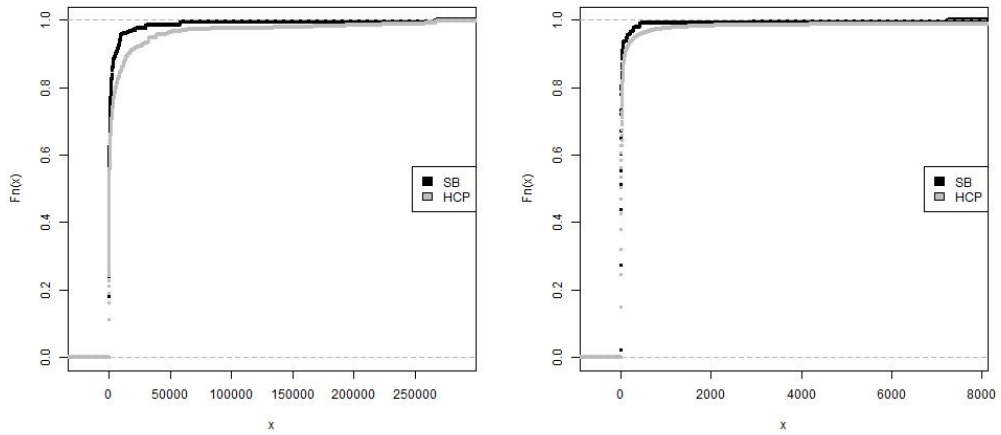


Figure 5.3: Number of references

inventors. Likewise, the experience of the applicants of a patent is the maximum number of patent applications that any of its applicants has filed.



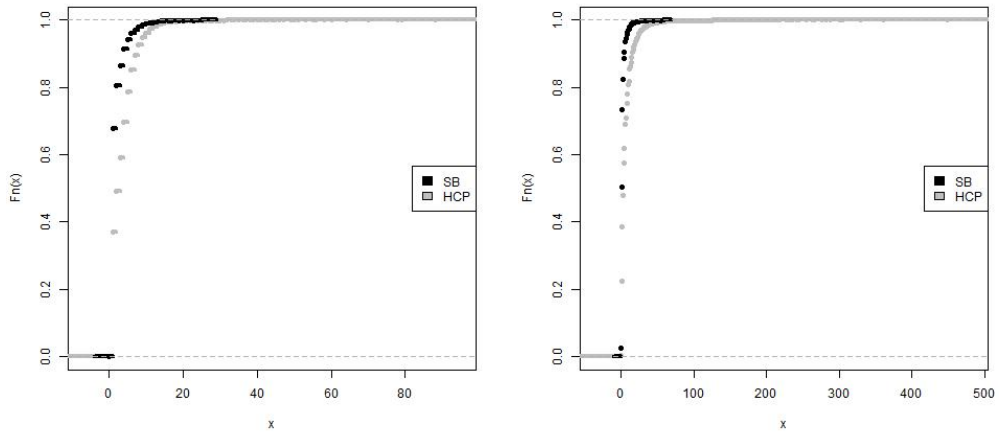
(a) Experience of the applicants *(b) Experience of the inventors*

Figure 5.4: Experience of the authors

Figure 5.4 shows the distribution of the experience of the authors, both for applicants (Figure 5.4*a*) and for inventors (Figure 5.4*b*). The experience of an applicant (inventor) is the number of patents she has applied for (invented), independently of the number of citations it received. The distribution of experience of the authors is extremely skewed. The vast majority of the authors have a limited

experience: around 20% of the applicants have only filed one patent application, and around 25% of the inventors have invented one or two other patents. On the other hand, there is one applicant who has filed over 300,000 patent applications, Matsushita Electric Inc Co Ltd. Four of those are in the SB group, and 2,145 of them are in the HCP group.

A Mann-Whitney test shows that the distribution of experience of the authors for SBs and HCP are different, both for applicants ($W = 21606.5, p - value < 0.001$) and for inventors ($W = 10927.5, p - value < 0.001$). The authors of SBs tend to be less experienced than the authors of HCP (the SB black line is above the HCP grey line). This could indicate that experienced authors are more able to recognize and diffuse a breakthrough invention. Nonetheless, very big players also produce some SBs.

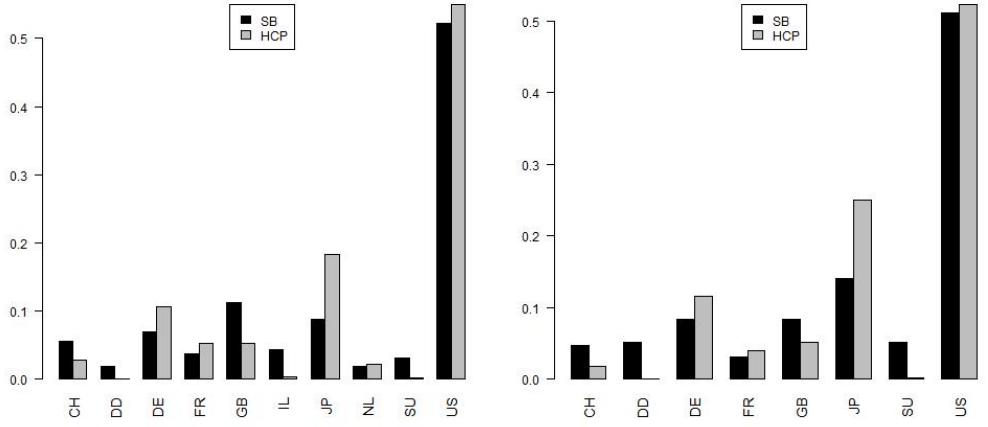


(a) Number of applicants *(b) Number of inventors*

Figure 5.5: Number of authors

Patents from big firms are likely to be the result of a big team work, so one would expect them to have more authors. Patent authors are separated in two groups: inventors and applicants. The applicant fills the application, and might be the inventor, her organization, or her patent attorney. Figure 5.5 shows the distribution of the number of applicants (Figure 5.5*a*) and number of inventors

(Figure 5.5*b*) per patent. A Mann-Whitney U test shows that the distribution of number of authors per patent differ from SBs to HCP, both for applicants ($W = 801964016, p-value < 0.001$) and for inventors ($W = 707462148, p-value < 0.001$). The number of authors per patent is lower for SBs (the black line is higher). This could mean that small teams are less capable to diffuse their inventions. This result could indicate that lone inventors are more likely to create breakthroughs outside the technological paradigm, but it could also indicate that lone inventors lack the social network of authors through which an invention gets diffused.



(a) Country of the applicants *(b) Country of the inventors*

Figure 5.6: Country of the authors

Moreover, every applicant or inventor has a country code that can be associated to the patent country of origin. Patents from a multinational firm with branches in many countries can be applied for in different patent organizations under the local branch of the firm (the inventor, nonetheless, is usually the same). In that case, every additional country of origin is added to the countries of the patent family, and every patent family is duplicated as many times as countries it has.

Figure 5.6 shows the distribution of countries of the SBs and HCP (only countries with a share higher than 0.005% are shown). A χ^2 goodness-of-fit test shows

that the distribution of countries is different for SBs and HCP, both for applicants ($X^2 = 836.8716, df = 174, p - value < 0.001$) and for inventors ($X^2 = 852.7427, df = 186, p - value < 0.001$). The result shows that there are less SBs originating in Japan than their share of HCP would suggest. The Japanese patent system has some peculiarities, compared to the rest of the systems, that could account for this difference. One the one hand, patents in Japan are more market-oriented than patents in other systems, so they are less likely to patent breakthroughs that will not be recognized as such at first. On the other hand, patents in Japan follow a "one patent, one claim" policy. This could indicate that highly cited patents in other patent systems are made of many important claims, that become highly cited patents in the Japanese system.

5.4.2 Regression analysis

In this section, the previous dimensions of comparison between SBs and HCP are introduced in a multivariate analysis through a regression of the probability that a breakthrough is recognized with delay. The dependent variable takes value 1 if the patent is a SB and 0 if it is a HCP in the control group. The population consists of 735,653 patents, of which 2,136 are SBs and 733,517 are HCP. The suitable regression model for this kind of data is a logistic regression.[6]

Two new variables are added: priority year, to account for changes in the sociotechnical structure; and number of technological classes (Nr IPC), to separate the transversal technologies that can be relevant to many domains. The effect of the country of the authors, although interesting, is left for further analyses. Independent variables are separated in two groups: the characteristics of the patents and the IPC. We run a regression for each group and another with all dependents

[6]SBs account for 0.29% of the whole population in our database, what statistics call a rare event. Methods of logistic regressions for rare events data have been recently developed by King and Zeng (2001), but they are only adequate when the data is a sample of the entire population. Since our database includes the whole population of patents with more than 12 citations, a regular logistic regression is more suitable for our study.

included (models 1-3). IPC classes are non-exclusive, so we run a second group of regressions (models 4 and 5) with a fractional count of the IPC classes that patents belong to (every patent is weighted down as many times as different IPC it is assigned to), with reference technological class Electricity (H). Results of the logistic regressions are summarized in Table 5.3.

The effect of the priority year is very consistent across specifications: later patents are less likely to be SBs than HCP. One important point is that this effect is not due to HCP being younger patents in our database, since all patents, both in the SB population and in the control group, have priority dates in 1967-1996. This can indicate a change in the incentives to patent, since later breakthroughs are immediately recognized as such while earlier patents might still contain some yet unrecognized breakthrough inventions.

The exploratory analysis showed that SBs tend to cite less references than HCP. Indeed, this result holds in the multivariate analysis. As already suggested, patents that cite few earlier references are less connected to already existing technologies. Thus, patents that are outside the existing technological paradigm are more likely to be recognized as breakthroughs with delay.

As suggested by the exploratory analysis, both the number of applicants and the number inventors have a negative effect on the probability to produce a SB. HCP produced by big teams are less likely to experience delayed recognition. This can indicate either that the SBs are developments at technological the frontier, that need of big teams, or that the social network of big teams is wider and facilitates the diffusion of a breakthrough more easily than the smaller network of lone inventors.

The effect of the authors experience, on the other hand, is less determinant, and it varies between specifications. Indeed, as already noted, big players can sometimes produce SBs as well.

	(1)	(2)	(3)	(4)	(5)
Priority year	−0.175***	−0.199***	−0.172***	−0.201***	−0.173***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Nr references	−0.059***		−0.062***		−0.063***
	(0.006)		(0.006)		(0.006)
Nr applicants	−0.176***		−0.174***		−0.178***
	(0.016)		(0.016)		(0.016)
Nr inventors	−0.023**		−0.024***		−0.025***
	(0.009)		(0.009)		(0.009)
Applicant exp	−0.000005***		−0.00000		−0.00000
	(0.00000)		(0.00000)		(0.00000)
Inventor exp	0.00000		−0.00000		−0.00000
	(0.00003)		(0.00003)		(0.00003)
Nr IPC	−0.250***				
	(0.049)				
IPC A		0.024	0.104	0.898***	0.822***
		(0.067)	(0.067)	(0.089)	(0.089)
IPC B		−0.374***	−0.237***	0.411***	0.431***
		(0.061)	(0.061)	(0.090)	(0.090)
IPC C		−0.280***	0.003	0.520***	0.706***
		(0.069)	(0.071)	(0.103)	(0.102)
IPC D		−0.210	0.086	0.561***	0.821***
		(0.129)	(0.129)	(0.178)	(0.176)
IPC E		−0.211**	−0.145	0.609***	0.538***
		(0.093)	(0.093)	(0.118)	(0.118)
IPC F		−0.396***	−0.238***	0.347***	0.389***
		(0.073)	(0.073)	(0.101)	(0.101)
IPC G		−0.849***	−0.733***	−0.219**	−0.216*
		(0.077)	(0.077)	(0.112)	(0.110)
IPC H		−0.702***	−0.580***		
		(0.081)	(0.081)		
IPC count	Full	Full	Full	Fractional	Fractional
Observations	735,653	735,653	735,653	735,653	735,653
Log Likelihood	-11,962.810	-12,142.570	-11,865.620	-12,168.460	-11,872.540

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5.3: Logistic regressions

Patents assigned to several technological classes are less likely to be unrecognized as breakthroughs, as signaled by the negative effect of *Nr IPC*. Since they have more potential applications, it is logical that some class or another will recognize them immediately. Moreover, IPCs are assigned by professional examiners. That SBs have less IPC classes might also indicate that examiners ignore their future impact and applications. Finally, patents in the technological classes G (physics) and H (electricity) are much less likely to remain unrecognized as breakthrough for a long time, compared to patents in other classes. This is evidenced by the fact that their coefficients are bigger than those of the other IPCs in models 2 and 3, and that the effect of all IPCs excepted G are positive respected the reference class H in models 4 and 5.

5.5 Conclusion

The study of breakthrough innovations is core for the analysis of technological development through different technological paradigms. Some breakthroughs, nonetheless, remain latent for a while until further technological development draws on them.

This study carries out an exploratory analysis of these technological SBs. To do so, it compares SBs to other breakthroughs in the same period across several dimensions, first unidimensionally, and then through a regression analysis. It points out some of the distinctive features of SBs.

First, SBs seem to be based on developments outside of the established technological trajectory, as signaled by the smaller number of references to prior technologies compared to HCP. This implies the need of big teams of inventors, indicating a rather complex or difficult technology. Indeed, SBs tend to have more inventors than HCP, providing evidence for this idea.

Second, SBs might slip past unnoticed in early periods of their lifetimes, both by their authors and by professional patent examiners. Inventors of SBs are less likely to get involved on the application process, indicating that they might not be aware of its future impact. Moreover, SBs tend to be assigned to less technological classes by the examiners. This suggest either that they contain very specific technologies, or that the examiners as well are unaware of the patent potential applications.

This study is not exempt of limitations. Mainly, this remains a purely exploratory analysis of the characteristics of breakthroughs with delayed recognition. It does not provide a clear theory as to why SBs appear, whether they result from a failure of the patent system or they are a natural consequence of technological development. This study only presents a first, descriptive approach to SBs and their features through the analysis of some empirical trends of their population.

The examination of SBs can be extended in several interesting directions. The exploratory analysis could be expanded by a social network analysis of the authors of SBs, to identify potential trends in their position on the overall network of authors. Moreover, a further step could be made in categorizing the citing patents of SBs. For example, SBs can be the result of migrating technologies, those that got an unexpected application in a field different from the one they were oriented to at first. Analyzing the citing patents can lead to an interesting further development: the study of the "awakening" patent, what is called the prince of the SB in scientific literature (van Raan, 2004). A network analysis both of this patent (in the network of citations) and of the authors (in the network of coauthorships) can point out whether the characteristics of the prince are crucial for the awakening of the SB. If not, it may be that SBs are a product of fashions or trends in the technological direction of industrial development.

CHAPTER 6

A NETWORK APPROACH TO THE DIVISION OF LABOR IN INDUSTRY

This chapter has been produced in collaboration with D Consoli. The PhD candidate has been the primary researcher of the work reported in this chapter and has been the main contributor in the following stages of research: theory, methodology, analysis, writing and presenting.

6.1 Introduction

The widespread availability of ever more refined data calls for methodological advances that match the potential for greater understanding of how economies are organized and of how their structures evolve. No doubt, recent advances in computerization have accelerated the proliferation of new research methods. Network analysis, in particular, has experienced significant progress in the study of the component subsystems of economies along the lines of the probabilistic analysis to graph theory first sketched by Erdos and Renyi (1959). One recent development in this field is the method proposed by Hidalgo and Hausmann (2009) to infer the nexus of capabilities underlying an economy by means of bipartite network of countries and products created from export data. This method has since

been adopted and adapted to a variety of data sources in with the goal of testing its robustness and usability.

The present chapter elaborates on their work to produce an empirical study of the occupational structures of 285 industrial sectors in the United States (US) over the period 2002-2012. Starting from the idea that jobs are the pathways through which useful knowledge is applied to production activities, we focus on the occupational configurations that characterize the prevailing division of labor within industry. In so doing we move on mere aggregate employment figures, and focus on the interdependence across occupations.

This study adds to existing literature by unveiling little explored conceptual and empirical aspects of industrial dynamics. For what concerns the former, we portray occupations as species that combine with each other in uniquely distinctive ways within sectors. This, we argue, is a rather original account of the relational and distributed nature of knowledge underpinning the division of labor. For what concerns the empirical strategy, our analysis shows that the organization of occupations generates bipartite networks that reflect the know-how that industries actually use. We show that the productivity differences across industries can be explained in large part by the diversity of occupations being combined, which indicates the extent of the division of labor present in an industry. We also find that the ubiquity of occupations across industries is positively related to productivity, which we understand as being due to their lower propensity to be externalized.

The structure of the chapter is as follows. Section 6.2 summarizes the associated existing literature. Section 6.3 applies the method of reflections to our dataset, and Section 6.4 provides some concluding thoughts.

6.2 Literature review

6.2.1 Industry dynamics

This section focuses on new ways to measure division of labour across industries and on how the latter affects productivity. There is an established literature about interindustry differences in productivity that dates back to Nelson and Winter (1977). The gist of this argument is that regardless of the measurement method or unit of analysis, technological progress tends to engender significantly different responses across industries. Various explanations have emerged as on why. One focuses on differences in industry R&D intensity and in firm size and market structure, though the evidence in support is rather weak (Cohen and Levin, 1989; Klevorick et al., 1995). A second line of inquiry calls attention to the role of demand in determining the extent of the market and therefore the level of innovative activity (Schmookler, 1966). Such a conjecture rests on the assumption that firms exploit a pool of generic and ready-to-use pool of scientific knowledge that allows them to respond rapidly to increases in demand. As Rosenberg (1976) and Walsh (1984) showed, however, the type of knowledge that spurs innovation and productivity growth tends to be contextual and driven by trial-and-error experimentation. Yet another approach ascribes productivity differences across industries to a combination of technological opportunity and the ability to appropriate returns from innovation. In particular, the former shapes R&D productivity while the latter determines the portion of the returns from investments that the innovator can retain. Appropriability conditions have been debated at length, especially in industrial economics (e.g. Mansfield et al., 1981; Spence, 1984), while the notion of technological opportunity has become the staple of evolutionary approaches based on the Schumpeterian distinction between creative destruction, whereby new firms enjoy productivity increases due to rad-

ical innovations, and creative accumulation, wherein established firms generate incremental innovation. Breschi et al. (2000) propose a synthesis of these issues in a seminal article that establishes the idea of technological regime as a combination of opportunity, appropriability, cumulativeness conditions and properties of the knowledge base (specific vs. generic) underlying differences in capacity to innovate across industries.

It is interesting to observe that while technological and institutional aspects of the division of labor within and across industries have been elaborated in depth (see Malerba, 2005, for a review), the notion of knowledge base has somewhat remained in the background. What do we know about the forms of knowledge that make up industry? Do different forms of organizing productive knowledge via the division of labor yield differential productivity performance? This section focuses on these questions, and elaborates a new method to characterize the complexity of the knowledge base by measuring both the diversity of occupations used in an industry and how ubiquitous these are across industries.

6.2.2 The method of reflections

This study is based on the method of reflections of Hidalgo and Hausmann (2009). The method of reflections uses two primary measures: the richness[1] of an industry (i.e. the number of occupations its workforce uses with RCA) and the ubiquity of an occupation (i.e. the number of industries that use that occupation with RCA). It then calculates jointly and iteratively the average of the preceding measures. For instance, in the second step it calculates the average ubiquity of the occupations that a sector uses and the average richness of the sectors that use an occupation. The intuition is that a sector that exhibits higher

[1] In studies of country production, this variable is named "diversification" (Caldarelli et al., 2012; Cimini et al., 2014; Felipe et al., 2012; Hidalgo and Hausmann, 2009). Diversification, nonetheless, is traditionally associated to the diversity in production rather than the diversity of skills, or occupations used. Thus, we have preferred to follow the terminology in Baudena et al. (2015), and call this measure "richness".

richness (i.e. its workforce uses more occupations) is "more complex" than a sector that uses less occupations. Likewise, occupations that are employed by less (more) sectors are more specialized (ubiquitous) and thus are more (less) complex. These measures offer a compact yet revealing indication of the complexity underlying the organization of a sector.

$$\begin{aligned} k_i^{(0)} &= \sum_{o=1}^{N_o} M_{io} \\ k_o^{(0)} &= \sum_{i=1}^{N_i} M_{io} \end{aligned} \tag{6.1}$$

The method is based on two variables, $k_i^{(n)}$ and $k_o^{(n)}$, for industries $i \in \{1 \dots N_i\}$ and occupations $o \in \{1 \dots N_o\}$, where N_i is the number of industries and N_o is the number of occupations. The first pair, $k_i^{(0)}$ and $k_o^{(0)}$ (Equation 6.1), are the sum of the rows and columns, respectively, of $M = (M_{io})_{i,o}$, the adjacency matrix of the bipartite network of industries and occupations. Thus, $k_i^{(0)}$ measures the richness of an industry, or how many occupations industry i uses. Likewise, $k_o^{(0)}$ measures the ubiquity of an occupation, or how many industries use occupation o. The rest of the variables are defined iteratively (Equation 6.2).

$$\begin{aligned} k_i^{(n)} &= \frac{1}{k_i^{(0)}} \sum_{o=1}^{N_o} M_{io} k_o^{(n-1)} \\ k_o^{(n)} &= \frac{1}{k_o^{(0)}} \sum_{i=1}^{N_i} M_{io} k_i^{(n-1)} \end{aligned} \tag{6.2}$$

The method of reflections has been applied since its introduction to several studies on economic complexity. Felipe et al. (2012) compared its results to different measures of technological capability using export data of 5107 products and 124 countries. They found that the variables for country complexity where

highly correlated to other existing measures such as the technology achievement index (Desai et al., 2002) or the indicator of technological capabilities (Archibugi and Coco, 2005). Caldarelli et al. (2012) used the method of reflections to build a product taxonomy based on the activity of countries, and analyzed in depth the mathematical implications of the methodology. More recently, Cimini et al. (2014) have applied the method of reflections to a different set of data to study the production of scientific papers by countries.

This method has also been exported to the field of ecology by Baudena et al. (2015). In their study, they ascertain the validity of the variables produced by the method of reflection to study the species richness of a site in relation to environmental measures such as precipitations or temperature. Throughout this study, we will adopt their labeling of the first variable of the method of reflection, originally called "diversification". While diversification made sense for country data on exports, we find that richness of occupations condenses better the idea of number of occupations that an industry uses.

So far, all applications of the method of reflections have included a geographical dimension. We will extend this approach with a further step in abstraction, in the same line of previous work by Neffke and Henning (2013), by exploring the knowledge base of industries rather than geographical regions, in terms of the occupations present in an industry.

An important feature of the method of reflections is that odd and even variables have different meanings. The variables $k_i^{(2n+2)}$ of the even iterations of industries are a refined version of the previous even variables $k_i^{(2n)}$, and they measure the richness of industries. Likewise, variables $k_o^{(2n+2)}$ of the even iterations of occupations are a refined version of $k_o^{(2n)}$, and they measure the ubiquity of occupations. On the other hand, the meaning of the odd variables of industries and occupations is more difficult to put into words. The first iteration can be explained in intuitive terms: $k_i^{(1)}$ is the mean ubiquity of the occupations that in-

dustry i uses, and $k_o^{(1)}$ is the mean richness of the industries that use occupation o. The rest of the iterations can be interpreted as increasingly refined measures of the complexity of an industry and an occupation: $k_i^{(2n+1)}$ is the mean value of $k_o^{(2n)}$ of the occupations that industry i uses, and $k_o^{(1)}$ is the mean value of $k_i^{(2n)}$ of the industries that use occupation o.

Several studies have pointed out some conceptual and practical flaws of the method of reflections (Baudena et al., 2015; Cristelli et al., 2013; Tacchella et al., 2012, 2013). First of all, variables change their meaning in odd and even iterations, which produces two disparate sets of variables from a single formula. This is a conceptual problem when interpreting the meaning of these variables, since they need to be presented in pairs of odd and even, and interpreted separately. One cannot anticipate the meaning of $k^{(n)}$ since it will differ for odd and even iterations. An iteration of a variable should refine the measure under investigation instead of completely switching its definition.

This shortcoming has an additional problem that has not yet been pointed out by the literature. In the original dataset of countries and products, very complex countries were defined as those that exported very complex products. Very complex products were defined as those that where exported by few countries, but only those that exported many other products. This way, there was a definite implication that a country needed a very specific set of capabilities in order to produce and export those products. Thus, a complex product scored low values in the even iterations $k_p^{(2n)}$ (here p stands for product), and high values of the odd iterations $k_p^{(2n)}$. Likewise, a country would get a higher complexity score if it produced these products, so it would get high values of the odd iterations, constructed from the previous even iteration of products, and low values of the even iterations. Nonetheless, every iteration is constructed as a mean of the previous one, and Hidalgo and Hausmann (2009) found a strong nestedness in the production structure of countries: countries that produced specific products tended

to produce also ubiquitous ones. Thus, countries that export products with low $k_p^{(2n)}$ would not necessarily score low on the following odd iteration since high values increase a mean to a wider extent that low values decrease it. Their analysis did manage to separate three countries, Singapore, Chile and Pakistan, with similar values of $k_c^{(0)}$ (here c stands for country) but very different capabilities and, indeed, very different $k_c^{(18)}$.

Another limitation of the method is that, as the iterations go towards convergence, the background noise is eliminated, but some information is also shrunk. Indeed, Tacchella et al. (2013) found that country variables were correlated to country capabilities only for early iterations. As the variables were refined, this correlation was lost. This effect can be caused by the differences in meaning of low an high values of the variables for the different iterations.

Many of these flaws can be solved by using only the first two iterations of the method, which are also the most intuitive and easy to put into words. Using only the first iterations means that we loose most of the information regarding the inherent complexity. Nonetheless, using together the first odd and even iterations mends this deficiency by supplying both how many occupations an industry uses, and how many industries use these occupations.

For the descriptive analysis of industries we will only use the first two iterations, $k_i^{(0)}$ and $k_i^{(1)}$, since they are the ones that we can explain in intuitive terms. Later in the empirical analysis, we will compare the results obtained with these two first variables and the refined iterations.

6.3 Analysis

The empirical analysis is based on the Occupational Information Network (O*NET) electronic database of the U.S. Department of Labour (DOL), a source of specific information on the task and skill requirement of more than 900 occupations.

For the purposes of this chapter we use on information concerning the physical and cognitive abilities that are required to carry out job tasks as provided by trained occupational analysts, job incumbents and occupational experts (National Research Council, 2010). In broad terms, the catalogue of skill encompasses information on basic skills (e.g. reading, writing and listening) and on cross-functional skills (e.g. problem-solving, technical maintenance, social skills, resource management skills, etc). We additionally match O*NET data with information on employment from the US Bureau of Labor Statistics by using the Standard Occupational Classification (SOC). This yields a panel dataset of 285 sectors (4-digit NAICS) with detailed information on occupation-specific skills for the period 2002-2012.

O*NET, the Occupational Information Network, is a database of worker attributes and job characteristics maintained by the U.S. Department of Labor (DOL) and the National Center for O*NET Development, through its contractor Research Triangle Institute. It is the replacement for the Dictionary of Occupational Titles (DOT) and the primary source of occupational information for the US labour market. Data Collection is carried out in two steps: (1) identification of a random sample of businesses expected to employ workers in the targeted occupations, and (2) selection of a random sample of workers in those occupations within those businesses. New data are collected by means of a survey circulated among job incumbents. Occupations in O*NET are defined according to the criteria of the Standard Occupational Classification (SOC) system. Data Collection provides descriptive ratings based on the questionnaire covering various aspects of the occupation: Worker Characteristics, Worker Requirements, Experience Requirements, Occupation Requirements, Occupational Characteristics, and Occupation-Specific Information. In addition to the questionnaires completed by workers and occupation experts, additional ratings are provided by occupation analysts. Responses from all three sources -workers, occupation

experts, and occupation analysts- are used to provide complete information for each occupation.

6.3.1 The bipartite industries-occupations network

Building on the premises laid out in the preceding subsection, we propose that industry is an ensemble of occupations. From this it follows that each occupation exists in its own right - in the sense of being ontologically identifiable as well as being an observable unit of analysis - but, also, that the totality of occupations that are employed within an industry have full economic meaning when considered as an ecosystem, that is, when their interaction is explicitly recognized. This is operationalized here by means of network analytical techniques that allow to account for the mutual relation between occupations.

We focus on bipartite network connecting two mutually exclusive sets, the set of industries and the set of occupations. The structure of the bipartite network provides information on the complexity of industries based on the set of occupations the use, and also on the complexity of occupations, based on the industries that use them. Several methods have been proposed to retrieve this inherent complexity. Hidalgo and Hausmann (2009) developed the method of reflections, based on export data, to find out the complexity of countries and products. The intuition behind their method is that complex products are more profitable and attractive, so many countries would want to produce them, but only those with complex capabilities can. Thus, complex products have to be among those produced by only a small set of countries. Nonetheless, very basic products are also produced by a small set of countries, those with a very basic set of capabilities that cannot produce the more complex, more attractive ones. In order to separate the products that are produced by few countries because they are complex from those that are produced by few countries because they are not as profitable, they suggest to check the level of richness of the countries that produce them.

If a country has complex capabilities, it will be likely that they produce many more products than if it has basic capabilities. Thus, the most complex products are likely to be those produced by a small number of diversified countries, and the most basic products are likely to be those produced by a small set of non-diversified countries.

We represent the data as a binary bipartite network, in which nodes can be separated into two groups, industries and occupations, such that links only connect nodes in different partitions, i.e. between industries and occupations. The bipartite adjacency matrix is $M = (M_{io})_{i,o}$, and its elements M_{io} take value 1 if a link exists between industry i and occupation o, and 0 otherwise. A connection between a sector and an occupation signifies that the former utilizes the latter. In order to distinguish which occupations an industry uses, we follow previous works (Caldarelli et al., 2012; Hausmann and Hidalgo, 2011; Hidalgo and Hausmann, 2009) and define that an industry i uses occupation o if the revealed comparative advantage RCA_{io} (as defined in Equation 6.3) is higher than a specific threshold. We take that threshold to be 1 as in the original paper by Hidalgo and Hausmann (2009) and on following works (e.g. Cristelli et al., 2013; Felipe et al., 2012; Hausmann and Hidalgo, 2011).

$$RCA_{io} = \frac{X_{io}}{\sum_{o'} X_{io'}} / \frac{\sum_{i'} X_{i'o}}{\sum_{i',o'} X_{i'o'}} \tag{6.3}$$

6.3.2 Richness of industries

The method of reflections has never been applied to a database such as ours, to analyze the network structure of industries' knowledge base. Thus, we start with an exploratory analysis of the two first iterations of the variables for industries ($k_i^{(0)}$ and $k_i^{(1)}$) and for occupations ($k_o^{(0)}$ and $k_o^{(1)}$). Variable $k_i^{(0)}$ is the richness of industry i and $k_o^{(0)}$ is the ubiquity of occupation o: $k_i^{(0)}$ is the number of occupations that industry i employs with $RCA > 1$, and $k_o^{(0)}$ is the number of industries that

use occupation o with $RCA > 1$. On the other hand, the interpretation of $k_i^{(1)}$ and $k_o^{(1)}$ requires some more thought: $k_i^{(1)}$ is the mean ubiquity of those occupations that industry i uses, while $k_o^{(1)}$ is the mean richness of those industries that use occupation o.

The second step consists of exploring the richness $k_i^{(0)}$ of industries. For the sake of simplicity, we will aggregate industries in their 2-digit NAICS group. For example, the group 61, Educational services, contains 61-1400, Business Schools and Computer and Management Training, as well as 61-1700, Educational Support Services. The average richness is computed by aggregating all industries in a same 2-digit NAICS code and computing the mean of the number of occupations they use, $k_i^{(0)}$. The remaining analyses in the chapter consider the disaggregated 4-digit NAICS classification. Table 6.1 shows the aggregated industries in our database ordered by their mean richness.

Here we observe that the macro-sector Management of Companies and Enterprises is an outlier with a average $k_i^{(0)}$ (143) that is twice as large as the followers'. The second block of sectors (average $k_i^{(0)}$ between 76.4 and 72) includes Professional, Scientific and Technical Services, Utilities and Educational Services. Further down, manufacturing industries, Wholesale Trade and Mining activities make up the third group (average $k_i^{(0)}$ between 59.6 and 58). At the bottom of the table, the less diversified sectors are Transportation & Warehousing, Accommodation services, Agriculture and Retail Trade. To reiterate, these sectors are arranged by the heterogeneity of their occupational structure, that is, by the diversity of competences that characterizes the population of occupations within them.

We can elucidate additional interesting features of this ranking by considering the distribution of the sub 4-digit NAICS sectors included in these macro-aggregations. This seems relevant in consideration of the sheer diversity in the scale of sector-populations at hand, whereby Management of Companies is a

2-dig NAICS	Industries	Average richness
55	Management of Companies and Enterprises	143
54	Professional, Scientific & Technical Services	76.4
22	Utilities	73
61	Educational Services	72.3
31-33	Manufacturing	59.6
42	Wholesale Trade	59.3
21	Mining, Quarrying, & Oil and Gas Extraction	58
62	Health Care & Social Assistance	53.3
56	Admin Support and Waste Manag & Remediation	51.5
52	Finance & Insurance	51.4
51	Information	49.9
23	Construction	49.6
81	Other Services (except Public Administration)	47.6
71	Arts, Entertainment, and Recreation	47.6
53	Real Estate and Rental & Leasing	40.6
48-49	Transportation & Warehousing	30.8
72	Accommodation and Food Services	30.6
11	Agriculture, Forestry, Fishing and Hunting	28.3
44-45	Retail Trade	26.3

Table 6.1: Aggregated 2-digit NAICS ordered by their average richness

unique bloc as opposed to manufacturing which stands at the opposite extreme (86 sub-sectors), or to other comparable high-skill service sectors such as Professional Services (9), Financial & Insurance (11) or Real Estate (8).

The right-hand side of Table 6.1 shows that lower digit sectors within Professional, Scientific & Technical Services, Educational Services, Manufacturing, Health Care & Social Assistance, Waste Management & Remediation Services and Other Services are more likely to populate respectively the highest and the intermediate terciles. A second block includes aggregates whose component sectors are more likely to be first in the intermediate tercile and then in the third: Finance & Insurance, Information, Arts, Entertainment & Recreation, Mining, Wholesale

Trade and Construction. An interesting difference within this group is that the distributions of the first three are symmetric whereas the remaining ones exhibit very low probabilities in the lowest tercile. Yet another striking feature of this group is the similarities within each sub component, considering that Finance & Insurance, Information and Arts, Entertainment & Recreation are all information intensive activities whereas the remaining three encompass mostly manual type of work tasks. At the opposite end of this rank are groups whose distribution is tilted towards the bottom tercile, namely Retail Trade, Accommodation and Food Services, Transportation and Warehousing, Agriculture and Real Estate. Again, looking closer we find that the sectoral distributions of the first three are biased towards low richness compared to Agriculture and Real Estate. Lastly, Utilities stands out as the only macro group that exhibits polarization in the distribution of its integrating 4-digit industries.

6.3.3 Ubiquity of occupations

As a descriptive approximation to the use of these variables in our database, we start by analyzing the ubiquity of occupations and the richness of industries.

We first explore the ubiquity of occupations in the 2-digit SOC classification of occupations. In order to do so, we aggregate all the 4-digit occupations contained in a same 2-digit class. For example, Management occupations (2-digit SOC 11) includes Chief Executives (11-1011), Natural Sciences Managers (11-9121), etc. Table 6.2 shows the 2-digit classification ordered by mean ubiquity.

Three of the four most ubiquitous occupations are high-skilled professions e.g. Management, Business and Financial Operation professionals and Computer and Mathematical specialists. These occupations are normally associated with intensive use of broad-encompassing analytical and interactive skills, such as devise strategies, personal interaction or problem-solving. Therein the only

2-dig SOC	Occupation Descriptor	Average ubiquity
11	Management	58.65
43	Office and Administrative Support	47.98
13	Business and Financial Operations	40.95
15	Computer and Mathematical	38.57
41	Sales and Related	30.68
17	Architecture and Engineering	28.04
53	Transportation and Material Moving	27.58
51	Production	25.00
49	Installation, Maintenance, and Repair	24.62
27	Arts, Design, Entertainment, Sports, and Media	17.94
37	Building and Grounds Cleaning and Maintenance	17.00
21	Community and Social Service	14.62
19	Life, Physical, and Social Science	12.19
35	Food Preparation and Serving Related	11.94
29	Healthcare Practitioners and Technical	9.78
47	Construction and Extraction	9.36
31	Healthcare Support	8.86
39	Personal Care and Service	8.61
23	Legal	8.33
45	Farming, Fishing, and Forestry	7.20
33	Protective Service	7.00
25	Education, Training, and Library	4.40

Table 6.2: Aggregated 2-digit occupations ordered by their average ubiquity

exception are Office and Administrative Support jobs which, according to the established literature, is often considered a mid-skill occupation.

On the opposite side of the table, the four less ubiquitous occupations are Legal, Farming, Fishing and Forestry, Protective Service and Education, Training and Library.

This table also brings to the fore important qualitative characteristics of the occupations under analysis, in particular the relation between the generality of an occupation versus the propensity to externalize the main tasks. For exam-

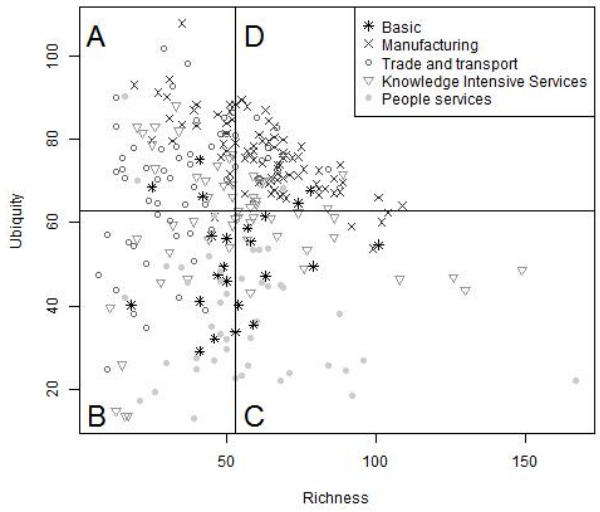


Figure 6.1: Industries plotted in their richness-ubiquity space, $k_i^{(0)}$ and $k_i^{(1)}$. The x-axis is the richness, or the number of occupations that industries use. The y-axis is the mean ubiquity, the number of industries that use those same occupations.

ple, management (SOC 11), computer (SOC 15) and sales (SOC 41) have high ubiquity while cleaning services (SOC 37), legal services (SOC 23) and protective services (SOC 33) have low ubiquity. All these six tasks are arguably essential and widespread across all industries, but the former requires in-depth knowledge of the daily running of a firm whereas the latter can be (and often is) outsourced to specialized cleaning, legal and protection firms. This illustrates the extent to which the ubiquity of an occupation is inversely related to its tradability.

6.3.4 Nestedness of occupations

Figure 6.1 depicts industries depending on their richness $k_i^{(0)}$ (horizontal axis) and the ubiquity of their occupations $k_i^{(1)}$ (vertical axis). The figure is divided in four quadrants,defined by the mean of both measures across all industries. Industries in the upper-left quadrant A have a higher than average ubiquity and lower than average richness.

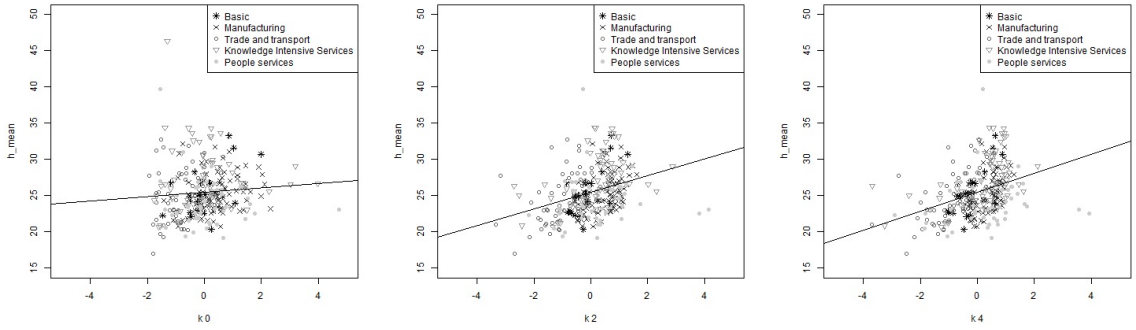
Richness and ubiquity exhibit a negative relation: industries with high richness, that is to say, that use many occupations, tend to score lower on ubiquity, that is to say, their occupations are used by a low number of other industries. This is an indicative of nestedness of the occupations: there is a set of specific occupations that are only used by a few industries, while there is another set of general occupations used by many industries.

According to the original definition by Hidalgo and Hausmann (2009), the most complex industries are the ones in quadrant C: those that use specific occupations with a low ubiquity, but use many other occupations, signaling that their set of capabilities is wide. Indeed, quadrant C is mostly occupied by grey dots, industries pertaining to People Services. Moreover, most Manufacturing industries (black crosses) are situated in the upper quadrants A and D, meaning that the occupations they use are employed by many other industries as well (a potential sign of low complexity).

In the following section we will explore how richness of industries in terms of occupations and the ubiquity of the occupations they employ relate to productivity. In the following, we will consider both the richness of industries in terms of occupations and the ubiquity of the occupations across industries as two measures of complexity. So, instead of using the term complex as Hidalgo and Hausmann (2009) used it to indicate industries that are present in only few and highly diverse countries, we use the term complex to indicate the richness of industries and the ubiquity of occupations.

6.3.5 Richness and productivity

Here we seek to establish empirical regularities between structural characteristics of the sector-occupation networks and performance indicators. Does a more complex knowledge base ensure higher productivity?



(a) Productivity vs $k_i^{(0)}$ normalized

(b) Productivity vs $k_i^{(2)}$ normalized

(c) Productivity vs $k_i^{(4)}$ normalized

Figure 6.2: Mean productivity of industries as a function of the first three measures of richness, normalized by their mean and standard deviation ($\frac{k_i^{(2n)} - mean_{i'}(k_{i'}^{(2n)})}{stdev_{i'}(k_{i'}^{(2n)})}$)

In order to fix ideas let us plot the distribution of sectors arranged by richness measures against their average productivity over the period 2002-2012. Figure 6.2 shows the average productivity of industries versus $k_i^{(0)}$, $k_i^{(2)}$ and $k_i^{(4)}$, normalized to facilitate the comparison. As commented before, the first measure of richness, the number of occupations that an industry uses, is not a good measure of complexity. Indeed, $k_i^{(0)}$ does not distinguish between specialized industries that use the occupations that require very low capabilities, so other industries outsource them, from those that use occupations that require very high capabilities, so not all industries can use them.

Indeed, the productivity of an industry and the number of occupations it uses are not correlated (Figure 6.2). Nonetheless, further iterations of the richness variable ($k_i^{(2)}$ and $k_i^{(4)}$) are positively correlated with an industry productivity. Since these iterations are refined measures of complexity, this signals that the more complex industries are also the more productive.

6.3.6 Productivity and complexity

The next step in our analysis consists in an empirical exploration of the correspondence between the skill makeup of sectors (as per the previous subsection) and a performance measure. For what concerns the latter we use aggregate (industry-level) labor productivity computed as value added per worker at the four-digit NAICS (Source: US BLS). Following the work by Hausmann and Hidalgo (2011); Hidalgo and Hausmann (2009), we estimate the relation between productivity ($h_i(t)$) and the first two measures of complexity, richness ($k_i^{(0)}$) and ubiquity ($k_i^{(1)}$), controlling for the lagged productivity ($h_i(t-1)$). Such an analysis allows to determine whether and to what extent structural properties of the occupation-sector networks explain sectoral productivity. Estimating Equation 6.4 with a Blundell-Bond dynamic panel data analysis (Blundell and Bond, 2000) allows to account for unobserved heterogeneity and endogeneity of the variables from the method of reflections. It also permits to introduce the lagged dependent variable $h_i(t-1)$, and it is the best suited method when the panel dataset has a short time dimension (from 2002 to 2012) but a larger number of industries (285). The results of the regression analysis are shown in Table 6.3. We control for the capital of industries (Office machinery and Infrastructures) as an instrument.

$$h_i(t) = a + b_1 \cdot h_i(t-1) + b_2 \cdot k_i^{(0)}(t) + b_3 \cdot k_i^{(1)}(t) \tag{6.4}$$

This regression analysis highlights several results. First of all, the effect of both the occupational richness of industries and the ubiquity of their occupations are positive and significant, and this effect is not affected by the use of capital and the number of firms as controls (models 1 and 2). Industries are more

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Lagged productivity	0.6537***	0.6531***	0.5487***	0.6591***	0.6645***	0.6655***	0.6660***
	(0.079)	(0.077)	(0.092)	(0.068)	(0.067)	(0.067)	(0.068)
$k_i^{(0)}$	0.0106***	0.0111***	0.0504*				
	(0.003)	(0.003)	(0.027)				
$k_i^{(1)}$	0.0123***	0.0127***	0.1326**				
	(0.004)	(0.004)	(0.066)				
Basic Industries			4.0732				
			(2.571)				
Transport & Trade			2.2819				
			(1.486)				
Knowl Intens Services			3.1584**				
			(1.573)				
People Services			5.8755*				
			(3.273)				
$k_i^{(2)}$				0.0539***			
				(0.017)			
$k_i^{(3)}$				0.0188***			
				(0.005)			
$k_i^{(4)}$					0.1421***		
					(0.040)		
$k_i^{(5)}$					0.0279***		
					(0.008)		
$k_i^{(6)}$						0.3337***	
						(0.086)	
$k_i^{(7)}$						0.0463***	
						(0.012)	
$k_i^{(8)}$							0.7353***
							(0.179)
$k_i^{(9)}$							0.0810***
							(0.020)
Capital	No	Yes	Yes	Yes	Yes	Yes	Yes
Tot Firms	No	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1911	1896	1896	1896	1896	1896	1896
N. of groups	273	271	271	271	271	271	271
AR2	-0.4004	-0.4084	0.904	-0.4333	-0.4233	-0.3934	-0.3648
AR2 crit. prob.	0.889	0.683	0.928	0.6648	0.6721	0.694	0.7153
Hansen J	23.2649	22.9114	25.8608	22.8885	23.3492	23.7232	24.0521
Hansen df	20	20	24	20	20	20	20
Hansen crit. prob.	0.276	0.2932	0.3603	0.2943	0.272	0.2547	0.2401
Instruments	31	33	41	33	33	33	33

Note: *p<0.1; **p<0.05; ***p<0.01

Table 6.3: Regression results

productive if they use many occupations, and if those occupations are used by many other industries. Thus, the most complex industries in Figure 6.1 (quadrant high richness and low ubiquity) are not the most productive, in contrast to the analysis by Hidalgo and Hausmann (2009) which was based on export data.

As already commented, the negative relation of $k_i^{(0)}$ and $k_i^{(1)}$ indicates that some occupations are widely used while others are very sector specific. The positive effect of ubiquity $k_i^{(1)}$ implies that occupational sector-specificity is not conducive to higher productivity. Moreover, it has also been noted that ubiquity of occupations was closely related to their tradability. One possible interpretation is that occupations that are used by many industries must necessarily be good for productivity, since otherwise industries would outsource them. Thus, those occupations that have survived in many sectors must be perceived as specially profitable.

This effect is also stable after controlling by the taxonomy that separates sectors in Manufacturing, Basic, Transport & Trade, KIBS, and People Services (model 3). Looking at the coefficient of the sector-group dummies we notice that the two groups of service industries, Knowledge Intensive Services (i.e. Telecomunications, FIRE and KIBS) and Personal Services (i.e. Social Services, Education, Cleaning) enjoy higher productivity relative to the reference group of Manufacturing industries. Indeed, the coefficients associated to richness and ubiquity increase after introducing these controls, indicating that there are some differences that are due to the group industries belong to, that were explained by these two variables in the previous models.

Finally, we study the effect of the more refined measures of richness and ubiquity for industries (models 4-7). The measure of richness $k_i^{(2n)}$ and the measure of ubiquity $k_i^{(2n+1)}$ can be refined to $k_i^{(2n+2)}$ and $k_i^{(2n+3)}$. According to Hidalgo and Hausmann (2009), the first two measures $k_i^{(0)}$ and $k_i^{(1)}$ would not reflect the full complexity of a sector. Indeed, with every new iteration the coefficient for the

measures of richness and ubiquity doubles from the previous one, signaling a noise reduction. Nonetheless, their effect remains qualitatively the same: positive and significant. The higher iterations, while informative and interesting, are not needed to testify the effect of richness and ubiquity.

It is unclear, though, which of richness or ubiquity is more important for productivity. On the one hand, models 1 and 2 show that the effects of both variables are comparable. After introducing the dummies for the sector groups (model 3), ubiquity becomes much more important, signaling that industries from a same group benefit more from using transversal occupations than from using many occupations. On the other hand, the refined variables (models 4-7) show that richness is much more important than ubiquity for productivity. That is to say, it seems more important to use many occupations than transversal ones. This may be due to how the variables are constructed from the previous ones, so that industries that use many transversal occupations score much higher than industries that use many specific or few transversal occupations.

6.4 Conclusion

The increasing amounts of economic data available call for new methods of data analysis. In this study we explore one recent methodology developed by Hidalgo and Hausmann (2009), the method of reflections, to analyze the underlying knowledge base of industries as evidenced by their occupational structure.

The method of reflections has been thoroughly studied and widely applied to economic data of countries to analyze some measure of their production (product export, scientific papers, etc). Here we take one further step of abstraction and analyze industries according to their occupational structure. Our analysis shows that the method of reflections is a proper methodology for this kind of data. First, because the rationale behind the method of reflections is sensible when

applied to this kind of economic data. Second, because the results it produces are reasonable and show coherence with the existing literature.

The study reveals several properties of the occupational structure of industries. First of all, occupations follow a nested pattern in their use by industries, as shown by the negative relation between richness and ubiquity. That is to say, industries that use few occupations use the more ubiquitous ones; and only industries that use many occupations use specific ones. This implies that industries only diversify their pool of occupations if they also use the more general ones.

Moreover, richness and ubiquity are reasonable predictors of the productivity of an industry. The most productive industries score high both on richness and ubiquity. That is to say, highly productive industries use many occupations, and the occupations associated with the highest productivity gains are transversal, used by many industries. Ubiquity of occupations is closely related to their tradability: the occupations that score higher in the ubiquity measure are those with a lower propensity to be externalized.

CHAPTER 7

Conclusion

This thesis presents five studies on the social dimension of knowledge, with a focus on its creation and diffusion processes. Broadly speaking, the creation and diffusion of knowledge are phenomena inherent to human society as a whole. In this sense, the results of the works in this thesis can be relevant and applicable in several social domains. However, the main focus of this work is on ideas rather than specific information and opinions. A working example that has been used through the chapters in this thesis is the dynamics of scientific ideas: how they are conceived, adopted and diffused within scientific environments.

The chapters have shed light on various aspects of knowledge creation and diffusion, which can be considered in the following sequence. First, researchers collaborate to produce scientific output (Chapter 2). Subsequently, this knowledge diffuses amid the scientific community (Chapter 3). In particular, researchers working on novel ideas and heterodox approaches strive to overturn established theories (Chapter 4). If they succeed, they create breakthrough advancements in

science, which sometimes experience delayed recognition. In such a case, they are called sleeping beauties (Chapter 5). Finally, Chapter 6 sheds empirical light on the division of labor in industries rather than in scientific knowledge production, as it can be mapped by data on a variety of professions. Though here the context is industrial, it does deal with the distributed nature of knowledge production (compare Chapter 2). All studies highlight that network-theoretical concepts can be applied fruitfully both theoretically and empirically as to understand knowledge creation and diffusion.

Let us further summarize the main findings of the various chapters. Chapter 2 builds a simulation model of knowledge creation in networks. Interacting agents create knowledge with a double feedback mechanism: their collaborations affect the amount of knowledge they create, which in turn affects their collaborative structure. The path dependent development of the simulations shows three types of coevolution patterns between the knowledge creation performance and number of collaborators: positive (the more collaborators, the more knowledge created), negative (too many collaborators are detrimental) and independent (performance does not depend on the number of collaborators). The feedback mechanisms between the process of collaboration and the process of knowledge creation originate some unexpected outcomes. Causality links between collaboration and performance become blurred. An overall positive coevolution scenario can be driven by the attractiveness of highly performing agents even if collaborations are not very productive in terms of knowledge creation. The main goal of this chapter has been to call attention to these feedback mechanisms between endogenous network formation and knowledge creation in collaborative settings.

Once produced, knowledge is diffused through a network. Chapter 3 studies the effect of social reinforcement in word-of-mouth diffusion of knowledge with a model of percolation in networks. It considers social reinforcement in different network structures and population types, focusing on the influence of strong ties

on the extent of diffusion, since their effect is not clear in the previous literature. The main result is that weak long ties work better for the diffusion of simple and self-evident ideas, while strong ties facilitate the diffusion of complex and controversial ideas in close-minded populations.

The diffusion of ideas in a network can lead to a shift of the established dominant paradigm. Chapter 4 describes scientific transitions that occur when alternative theories aim to become dominant. Every new theory exploits 'defections' from the original paradigm, so that social reinforcement towards alternative theories grows cumulatively. A new theory that arrives at the point of critical fragmentation (that is to say, once the pool of users of the original paradigm is shattered) absorbs all adopters of alternative theories and becomes dominant, replacing the initial paradigm.

One possible trigger for a paradigm shift is the introduction of a breakthrough development. Chapter 5 explores the delayed recognition of breakthrough knowledge with the identification and empirical analysis of "sleeping beauties" in technology. This extension of the concept of sleeping beauties from science to technology is a novel approach to acknowledge that not only scientific but also radical technological developments may fail to diffuse in early periods of their lifetime. Specifically, sleeping beauties tend to be developments outside the current technological trajectory. This is suggested by their relatively low number of references to prior art, and the relatively large size of their team of authors, as compared to other highly cited patents. On the other hand, sleeping beauties can be diffusion failures, as indicated by the lower experience of their authors and the fact that breakthroughs in the most codified technological domains are less likely to suffer from delayed recognition.

Finally, Chapter 6 studies the knowledge base of industries through the network structure of their occupational division of labor. It utilizes recent network analysis techniques to build novel indicators of the richness of industries and

the ubiquity of occupations. The negative relation between richness and ubiquity shows that occupations follow a nested pattern in the intensity of their use by industries. That is to say, only industries that use many occupations use the most specific ones, while most industries utilize the most ubiquitous occupations. Richness and ubiquity are also sensible predictors of the productivity of industries. Highly productive industries use many occupations. Moreover, the occupations that bring the highest productivity gain are the most ubiquitous, those with a lower propensity to be externalized. An obvious extension of this chapter is to repeat the exercise on the variety of professions in industries in the context of science by looking at the variety of professions present in different scientific disciplines.

Methodologically, we can conclude that recent advances in network science provide a wide array of concepts and models that can be used to understand the social dimension of knowledge. While the application of network theory in models of diffusion in social networks (Chapters 3, 4 and 5) goes back a long time, the application of network science to questions related to the production of knowledge is much more recent. In this sense, Chapter 2 and Chapter 6 are examples of innovative applications, while Chapters 3, 4 and 5 can be considered as further deepening the already rich diffusion research.

A number of policy implications can be identified for the results outlined above. Before going into these implications, it should be stressed that to apply network science to real policy questions, the models and empirical exercises in this thesis should be further qualified and specified to take into account the context. Hence, the policy implications that can be drawn from the various chapters are necessarily general, and call for more specific models and studies that may take one of the frameworks presented here as a core model.

The first general policy insight stems from Chapter 4 dealing with transitions. This chapter showed that transitions between paradigms - be them scientific or

technological paradigms - may benefit from being preceded by many trials. The reasons why variety may spur a transition is that varieties draw away consumers from the old paradigm and, if open to any new paradigm having abandoned the old paradigm, can be easily drawn into a new paradigm. In this process, every new variant fragments to agent base of the old paradigm making it increasingly vulnerable to a transition towards a new paradigm. This theoretical result is important, because it questions standard economic theory that posits that more variety would make a transition less likely as any new variant would have more difficulty to create a critical mass. In the spirit of van den Bergh (2008) who already argued that variety may support a transition process if varieties can be recombined, we showed that variety can also make a transition more likely if deviant users exert social pressure to adopt any new paradigm. The takeaway for policy is that fostering variety is not necessarily at odds at accelerating the transition process especially in contexts where innovative agents exploring different variants exert social pressure among them. It also provided a theoretical explanation that a specific policy leading to a new variant may at first look as a failure if diffusion is low, but may actually support the diffusion of a future technology.

More generally, the thesis can be linked to the Europe 2020 flagship initiative “Innovation Union”, in which the European Commission signals some of the vulnerabilities and opportunities that European regions and firms face in order to promote innovation with the aim of improving their future competitiveness. Some of the topics they address are closely related to the individual chapters and the overall motivation of this dissertation. As a means to further the double feedback between knowledge creation and diffusion, the Commission advocates to promote knowledge transfer and collaborative research, via specific strategies such as the promotion of open access for publicly funded research or the funding of transnational collaboration. Moreover, market fragmentation is identified as an unfavorable condition for innovation that must be tackled with policy action,

while our results suggest that fragmentation is, on the contrary, an opportunity for transitions. In this sense, this thesis can also deepen and enrich the policy debate in many directions.

Various avenues for future research can be envisaged. Specific suggestions for further research were already articulated at the end of the respective chapters. Here, we discuss some more general avenues.

The first three chapters of the thesis present studies based on simulation models. A natural extension of such work is an empirical validation of the models and their main conclusions. Models can be tested against longitudinal empirical data on collaborative knowledge production or diffusion, to verify the temporal patterns found in the simulations.

The chapters based on simulation and numerical observations can also be extended to include possible exogenous changes that would allow to test policies directed to affect the outcome of the process. Policy actions could be examined safely in a virtual environment. This could be implemented in the models to test the effect of changes in the parameters during the simulation. What is more, as already stressed, when models are applied in a specific policy context, some models features can be empirically determined by calibration (for example, the network structure among agents or individual preferences). This would make the simulation of policies both more realistic to the extent that the empirical data can be considered as reliable, and more feasible to carry out (since fewer combination of parameters have to be checked)

The diffusion models can also be used to test the effect of targeting particular nodes. Since social reinforcement builds with additional adopting neighbors, this could be used to see whether targeting nodes close to the hubs would be more effective that targeting the nodes of higher degree directly or some other set of nodes.

Moreover, the simulation chapters can be merged in a model of knowledge creation and diffusion, where the agents that create knowledge pertain to the network through which it is diffused. In such a case, the network would evolve endogenously between the knowledge creating agents. This model would be closer to the self-reinforcing cycle of knowledge creation and diffusion.

Furthermore, breakthroughs have been found to behave differently in science-based technological areas. An interesting extension would be to further explore this avenue with a sample of academic patents. Since patents with scientific inventors are more likely to be science-based, one would expect that academic patents are more likely to be sleeping beauties in the sense defined by Chapter 5. Identifying the causes and consequences of academic patents becoming sleeping beauties would allow detecting potential problems that scientists might face when patenting and commercializing their research. It would also point out possible solutions to improve the use of patents by universities.

Finally, in all chapters the focus has been on the social dimension of knowledge, while the frameworks proposed can in principle be extended to include other proximity dimensions as well (Boschma, 2005). One aspect in which our studies can be further specified in the cognitive dimension. This would extend the current models that refer to a single scientific community to contexts in which multiple scientific communities co-exist with different degrees of cognitive proximity. Another aspect that would deserve more attention in the future research is geography. By placing agents both in a social context (network proximity) and a geographical context (spatial proximity), more complex outcomes can be expected. For example, agents that are both socially and spatially proximate may be better able to send a signal or exert more social pressure than agents that are only socially proximate and spatially distant. The resulting network dynamics, then, will depend on the complex relations between network structure and agents' physical locations.

All these suggestions for applications and further research exemplify on the one hand that our studies have been stylized and should be interpreted as such while on the other hand that the framework proposed are generic and allow for many interesting applications and extensions to become relevant in various scientific disciplines as well in various empirical contexts.

Bibliography

Ahn, Y.-Y., Han, S., Kwak, H., Moon, S., and Jeong, H. (2007). Analysis of topological characteristics of huge online social networking services. In *WWW '07 Proceedings of the 16th international conference on World Wide Web*, 835–844, New York, NY, USA. ACM Press.

Ahrweiler, P., Pyka, A., and Gilbert, N. (2004). Simulating knowledge dynamics in innovation networks (SKIN). In Leombruni, R. and Richiardi, M., editors, *Industry and labor dynamics: the agent-based computational economics approach*. World Scientific Press, Singapore.

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

Ahuja, G. and Lampert, C. M. (2001). Entrepreneurship in the large corporation: a longitudinal study of how established firms create breakthrough inventions. *Strategic Management Journal*, 22 (6-7): 521–543.

Albert, M., Avery, D., Narin, F., and McAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research Policy*, 20 (3): 251–259.

Albert, R. and Barabasi, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74 (1): 47–97.

Alkemade, F. and Castaldi, C. (2005). Strategies for the diffusion of innovations on social networks. *Computational Economics*, 25 (1-2): 3–23.

Alkemade, F., Frenken, K., Hekkert, M. P., and Schwoon, M. (2009). A complex systems methodology to transition management. *Journal of Evolutionary Economics*, 19 (4): 527–543.

Aral, S., Muchnik, L., and Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences*, 106 (51): 21544–21549.

Archibugi, D. and Coco, A. (2005). Measuring technological capabilities at the country level: a survey and a menu for choice. *Research Policy*, 34 (2): 175–194.

Arthur, W. B. (1989). Competing technologies, increasing returns, and lock-in by historical events. *The Economic Journal*, 99 (394): 116–131.

Arthur, W. B. (2009). *The nature of technology: what it is and how it evolves*. Free Press, New York, NY, USA.

Azagra-Caro, J., Aznar-Marquez, J., and Blanco, J. M. (2008). Interactive vs. non-interactive knowledge production by faculty members. *Applied Economics*, 40 (10): 1289–1297.

Azagra-Caro, J. M., Yegros-Yegros, A., and Archontakis, F. (2006). What do university patent routes indicate at regional level? *Scientometrics*, 66 (1): 219–230.

Bakker, J., Verhoeven, D., Zhang, L., and Van Looy, B. (2016). Patent citation indicators: one size fits all? *Scientometrics*, 106 (1): 187–211.

Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, 519–528, Lyon, France. ACM Press.

Balland, P.-A., De Vaan, M., and Boschma, R. (2012). The dynamics of interfirm networks along the industry life cycle: the case of the global video game industry, 1987-2007. *Journal of Economic Geography*, 13 (5): 741–765.

Barabasi, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286 (5439): 509–512.

Barabasi, A., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., and Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311 (3-4): 590–614.

Barber, B. (1961). Resistance by scientists to scientific discovery: this source of resistance has yet to be given the scrutiny accorded religious and ideological sources. *Science*, 134 (3479): 596–602.

Baudena, M., Sanchez, A., Georg, C.-P., Ruiz-Benito, P., Rodriguez, M. A., Zavala, M. A., and Rietkerk, M. (2015). Revealing patterns of local species richness along environmental gradients with a novel network tool. *Scientific Reports*, 5: 11561.

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

Bell, G. G. (2005). Clusters, networks, and firm innovativeness. *Strategic Management Journal*, 26 (3): 287–295.

Blundell, R. and Bond, S. (2000). GMM estimation with persistent panel data: an application to production functions. *Econometric Reviews*, 19 (3): 321–340.

Bollen, J., Goncalves, B., Ruan, G., and Mao, H. (2011). Happiness is assortative in online social networks. *Artificial Life*, 17 (3): 237–251.

Borner, K., Maru, J. T., and Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences*, 101 (Supplement 1): 5266–5273.

Boschma, R. (2005). Proximity and innovation: a critical assessment. *Regional Studies*, 39 (1): 61–74.

Boschma, R. A. and ter Wal, A. L. J. (2007). Knowledge networks and innovative performance in an industrial district: the case of a footwear district in the South of Italy. *Industry & Innovation*, 14 (2): 177–199.

Bosque, G., Folch-Fortuny, A., Pico, J., Ferrer, A., and Elena, S. F. (2014). Topology analysis and visualization of Potyvirus protein-protein interaction network. *BMC Systems Biology*, 8 (1): 129.

Breschi, S., Malerba, F., and Orsenigo, L. (2000). Technological regimes and Schumpeterian patterns of innovation. *The Economic Journal*, 110 (463): 388–410.

Bruckner, E., Ebeling, W., Montano, M. A. J., and Scharnhorst, A. (1996). Nonlinear stochastic effects of substitution – an evolutionary approach. *Journal of Evolutionary Economics*, 6 (1): 1–30.

Burt, R. (1992). *Structural holes: the social structure of competition*. Harvard University Press, Cambridge, MA, USA.

Caldarelli, G., Cristelli, M., Gabrielli, A., Pietronero, L., Scala, A., and Tacchella, A. (2012). A network analysis of countries' export flows: firm grounds for the building blocks of the economy. *PLoS ONE*, 7 (10): e47278.

Campbell, A. (2013). Word-of-mouth communication and percolation in social networks. *American Economic Review*, 103 (6): 2466–2498.

Cantono, S. and Silverberg, G. (2009). A percolation model of eco-innovation diffusion: the relationship between diffusion, learning economies and subsidies. *Technological Forecasting and Social Change*, 76 (4): 487–496.

Cassiman, B. and Veugelers, R. (2006). In search of complementarity in innovation strategy: internal R&D and external knowledge acquisition. *Management Science*, 52 (1): 68–82.

Castaldi, C., Frenken, K., and Los, B. (2015). Related variety, unrelated variety and technological breakthroughs: an analysis of US state-level patenting. *Regional Studies*, 49 (5): 767–781.

Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, 329 (5996): 1194–1197.

Centola, D. (2011). An experimental study of homophily in the adoption of health behavior. *Science*, 334 (6060): 1269–1272.

Centola, D., Eguiluz, V. M., and Macy, M. W. (2007). Cascade dynamics of complex propagation. *Physica A: Statistical Mechanics and its Applications*, 374 (1): 449–456.

Centola, D. and Macy, M. (2007). Complex contagions and the weakness of long ties. *American Journal of Sociology*, 113 (3): 702–734.

Cimini, G., Gabrielli, A., and Sylos Labini, F. (2014). The scientific competitiveness of nations. *PLoS ONE*, 9 (12): e113470.

Cohen, W. M. and Levin, R. C. (1989). Empirical studies of innovation and market structure. In *Handbook of Industrial Organization*, volume 2, 1059–1107. Elsevier.

Cohen, W. M., Nelson, R. R., and Walsh, J. P. (2002). Links and impacts: the influence of public research on industrial R&D. *Management Science*, 48 (1): 1–23.

Cole, S. (1970). Professional standing and the reception of scientific discoveries. *American Journal of Sociology*, 76 (2): 286–306.

Cooke, P. and Wills, D. (1999). Small firms, social capital and the enhancement of business performance through innovation programmes. *Small Business Economics*, 13 (3): 219–234.

Cowan, R. and Jonard, N. (2003). The dynamics of collective invention. *Journal of Economic Behavior and Organization*, 52 (4): 513–532.

Cowan, R. and Jonard, N. (2004). Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control*, 28 (8): 1557–1575.

Cowan, R., Jonard, N., and Zimmermann, J.-B. (2004). On the creation of networks and knowledge. In Gallegati, P. M., Kirman, P. A. P., and Marsili, D. M.,

editors, *The Complex Dynamics of Economic Interaction*, number 531 in Lecture Notes in Economics and Mathematical Systems, 337–353. Springer Berlin Heidelberg, Berlin, Heidelberg.

Cowan, R., Jonard, N., and Zimmermann, J.-B. (2006). Evolving networks of inventors. *Journal of Evolutionary Economics*, 16 (1-2): 155–174.

Cowan, R., Jonard, N., and Zimmermann, J.-B. (2007). Bilateral collaboration and the emergence of innovation networks. *Management Science*, 53 (7): 1051–1067.

Cristelli, M., Gabrielli, A., Tacchella, A., Caldarelli, G., and Pietronero, L. (2013). Measuring the intangibles: a metrics for the economic complexity of countries and products. *PLoS ONE*, 8 (8): e70726.

Dakhli, M. and De Clercq, D. (2004). Human capital, social capital, and innovation: a multi-country study. *Entrepreneurship & Regional Development*, 16 (2): 107–128.

David, P. A. (1985). Clio and the economics of QWERTY. *The American Economic Review*, 75 (2): 332–337.

De Solla Price, D. (1965). Networks of scientific papers. *Science*, 149 (3683): 510–515.

Desai, M., Fukuda-Parr, S., Johansson, C., and Sagasti, F. (2002). Measuring the technology achievement of nations and the capacity to participate in the network age. *Journal of Human Development*, 3 (1): 95–122.

Dow, P. A., Adamic, L. A., and Friggeri, A. (2013). The anatomy of large facebook cascades. In *Seventh International AAAI Conference on Weblogs and Social Media*.

Erdos, P. and Renyi, A. (1959). On random graphs, I. *Publicationes Mathematicae*, 6: 290–297.

Fagerberg, J. and Verspagen, B. (2009). Innovation studies – the emerging structure of a new scientific field. *Research Policy*, 38 (2): 218–233.

Fagiolo, G. and Dosi, G. (2003). Exploitation, exploration and innovation in a model of endogenous growth with locally interacting agents. *Structural Change and Economic Dynamics*, 14 (3): 237–273.

Felipe, J., Kumar, U., Abdon, A., and Bacate, M. (2012). Product complexity and economic development. *Structural Change and Economic Dynamics*, 23 (1): 36–68.

Fleming, L. (2001). Recombinant uncertainty in technological search. *Management Science*, 47 (1): 117–132.

Fleming, L. and Frenken, K. (2007). The evolution of inventor networks in the Silicon Valley and Boston regions. *Advances in Complex Systems*, 10 (1): 53–71.

Fogli, A. and Veldkamp, L. (2012). Germs, social networks and growth. Technical Report w18470, National Bureau of Economic Research, Cambridge, MA, USA.

Foray, D. (2004). *Economics of knowledge*. MIT Press, Cambridge, MA, USA.

Frenken, K., Hekkert, M., and Godfroij, P. (2004). R\&D portfolios in environmentally friendly automotive propulsion: variety, competition and policy implications. *Technological Forecasting and Social Change*, 71 (5): 485–507.

Frenken, K., Izquierdo, L. R., and Zeppini, P. (2012). Branching innovation, recombinant innovation, and endogenous technological transitions. *Environmental Innovation and Societal Transitions*, 4: 25–35.

Frenken, K. and Verbart, O. (1998). Simulating paradigm shifts using a lock-in model. In Ahrweiler, P. and Gilbert, N., editors, *Computer Simulations in Science and Technology Studies*, 117–127. Springer Berlin Heidelberg.

Gardner, M. (1970). The fantastic combinations of John Conway's new solitaire game 'life'. *Scientific American*, 223: 120–123.

Geels, F. W. (2010). Ontologies, socio-technical transitions (to sustainability), and the multi-level perspective. *Research Policy*, 39 (4): 495–510.

Graf, H. (2006). *Networks in the innovation process: local and regional interactions*. Edward Elgar, Cheltenham, UK.

Granovetter, M. (1978). Threshold models of collective behavior. *American Journal of Sociology*, 83 (6): 1420–1443.

Grebel, T. (2012). Network evolution in basic science. *Journal of Evolutionary Economics*, 22 (3): 443–457.

Grimpe, C. and Kaiser, U. (2010). Balancing internal and external knowledge acquisition: the gains and pains from R&D outsourcing. *Journal of Management Studies*, 47 (8): 1483–1509.

Guler, I. and Nerkar, A. (2012). The impact of global and local cohesion on innovation in the pharmaceutical industry. *Strategic Management Journal*, 33 (5): 535–549.

Hall, B. H., Jaffe, A., and Trajtenberg, M. (2005). Market value and patent citations. *The RAND Journal of Economics*, 36 (1): 16–38.

Harhoff, D., Narin, F., Scherer, F. M., and Vopel, K. (1999). Citation frequency and the value of patented inventions. *Review of Economics and Statistics*, 81 (3): 511–515.

Harhoff, D., Scherer, F. M., and Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy*, 32 (8): 1343–1363.

Hausmann, R. and Hidalgo, C. A. (2011). The network structure of economic output. *Journal of Economic Growth*, 16 (4): 309–342.

Helbing, D. (2012). Agent-based modeling. In Helbing, D., editor, *Social Self-Organization*, 25–70. Springer Berlin Heidelberg.

Hethcote, H. W. (1989). Three basic epidemiological models. In Levin, S. A., Hallam, T. G., and Gross, L. J., editors, *Applied Mathematical Ecology*, volume 18, 119–144. Springer Berlin Heidelberg.

Hidalgo, C. A. and Hausmann, R. (2009). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences*, 106 (26): 10570–10575.

Hohnisch, M., Pittnauer, S., and Stauffer, D. (2008). A percolation-based model explaining delayed takeoff in new-product diffusion. *Industrial and Corporate Change*, 17 (5): 1001–1017.

Hughes, T. P. (2004). *American genesis: a century of invention and technological enthusiasm, 1870-1970*. University of Chicago Press, IL, USA.

Ibarra, H. (1993). Network centrality, power, and innovation involvement: determinants of technical and administrative roles. *Academy of Management Journal*, 36 (3): 471–501.

Jaffe, A. B., Trajtenberg, M., and Fogarty, M. S. (2000). Knowledge spillovers and patent citations: evidence from a survey of inventors. *American Economic Review*, 90 (2): 215–218.

Jewkes, J., Sawers, D., and Stillerman, R. (1969). *The sources of invention*. The Norton library N502. W. W. Norton, New York, NY, USA, 2d edition.

Kahneman, D. (2003). Maps of bounded rationality: psychology for behavioral economics. *American Economic Review*, 93 (5): 1449–1475.

Ke, Q., Ferrara, E., Radicchi, F., and Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences*, 112 (24): 7426–7431.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 115 (772): 700–721.

King, G. and Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9: 137–163.

Klevorick, A. K., Levin, R. C., Nelson, R. R., and Winter, S. G. (1995). On the sources and significance of interindustry differences in technological opportunities. *Research Policy*, 24 (2): 185–205.

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

Konig, M. D., Battiston, S., Napoletano, M., and Schweitzer, F. (2012). The efficiency and stability of R&D networks. *Games and Economic Behavior*, 75 (2): 694–713.

Kuhn, T. S. (1970). *The structure of scientific revolutions*, volume 2 of *International encyclopedia of unified science. Foundations of the unity of science*. University of Chicago Press, IL, USA.

Kuperman, M. and Abramson, G. (2001). Small world effect in an epidemiological model. *Physical Review Letters*, 86 (13): 2909–2912.

Kuperman, M. N. (2013). Invited review: epidemics on social networks. *Papers in Physics*, 5: 050003.

Lachance, C. and Lariviere, V. (2014). On the citation lifecycle of papers with delayed recognition. *Journal of Informetrics*, 8 (4): 863–872.

Laursen, K. and Salter, A. (2006). Open for innovation: the role of openness in explaining innovation performance among UK manufacturing firms. *Strategic Management Journal*, 27 (2): 131–150.

Lehmann, J., Goncalves, B., Ramasco, J. J., and Cattuto, C. (2012). Dynamical classes of collective attention in twitter. In *Proceedings of the 21st international conference on World Wide Web*, 251–260, Lyon, France. ACM Press.

Li, S., Yu, G., Zhang, X., and Zhang, W. (2014). Identifying princes of Sleeping Beauty – knowledge mapping in discovering princes. In *2014 International Conference on Management Science & Engineering (ICMSE)*, 912–918, Helsinki, Finland. IEEE.

Lissoni, F. (2005). The reaper and the scanner: indivisibility-led incremental innovations and the adoption of new technologies. *Cambridge Journal of Economics*, 29 (3): 359–379.

Llerena, P. and Ozman, M. (2013). Networks, irreversibility and knowledge creation. *Journal of Evolutionary Economics*, 23 (2): 431–453.

Lu, L., Chen, D.-B., and Zhou, T. (2011). The small world yields the most effective information spreading. *New Journal of Physics*, 13 (12): 123005.

Malerba, F. (2005). Sectoral systems: how and why innovation differs across sectors. In Fagerberg, J., Mowery, D., and Nelson, R., editors, *The Oxford Handbook of Innovation*, 380–406. Oxford University Press, UK.

Malerba, F., Nelson, R., Orsenigo, L., and Winter, S. (1999). 'History-friendly' models of industry evolution: the computer industry. *Industrial and Corporate Change*, 8 (1): 3–40.

Malva, A. D., Kelchtermans, S., Leten, B., and Veugelers, R. (2015). Basic science as a prescription for breakthrough inventions in the pharmaceutical industry. *The Journal of Technology Transfer*, 40 (4): 670–695.

Mansfield, E., Schwartz, M., and Wagner, S. (1981). Imitation costs and patents: an empirical study. *The Economic Journal*, 91 (364): 907–918.

Martinez-del Rio, J. and Cespedes-Lorente, J. (2013). Competitiveness and legitimation: the logic of companies going green in geographical clusters. *Journal of Business Ethics*, 120 (1): 131–146.

McFadyen, M. A. and Cannella, A. A. (2004). Social capital and knowledge creation: diminishing returns of the number and strength of exchange relationships. *Academy of Management Journal*, 47 (5): 735–746.

Melese, T., Lin, S. M., Chang, J. L., and Cohen, N. H. (2009). Open innovation networks between academia and industry: an imperative for breakthrough therapies. *Nature Medicine*, 15 (5): 502–507.

Molina-Morales, F. X. and Martinez-Fernandez, M. T. (2009). Too much love in the neighborhood can hurt: how an excess of intensity and trust in relationships may produce negative effects on firms. *Strategic Management Journal*, 30 (9): 1013–1023.

Moore, C. and Newman, M. E. J. (2000). Epidemics and percolation in small-world networks. *Physical Review E*, 61 (5): 5678–5682.

Mullen, B., Johnson, C., and Salas, E. (1991). Productivity loss in brainstorming groups: a meta-analytic integration. *Basic and Applied Social Psychology*, 12 (1): 3–23.

Neffke, F. and Henning, M. (2013). Skill relatedness and firm diversification. *Strategic Management Journal*, 34 (3): 297–316.

Nelson, R. R. and Winter, S. G. (1977). In search of a useful theory of innovation. In Stroetmann, K. A., editor, *Innovation, Economic Change and Technology Policies*, 215–245. Birkhauser Basel, Basel.

Nemet, G. F. (2009). Demand-pull, technology-push, and government-led incentives for non-incremental technical change. *Research Policy*, 38 (5): 700–709.

Newman, M. and Watts, D. (1999). Scaling and percolation in the small-world network model. *Physical Review E*, 60 (6): 7332–7342.

Newman, M. E. J. (2001). From the cover: the structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98 (2): 404–409.

Newman, M. E. J. (2002). Spread of epidemic disease on networks. *Physical Review E*, 66 (1): 016128.

Ozman, M. (2009). Inter-firm networks and innovation: a survey of literature. *Economics of Innovation and New Technology*, 18 (1): 39–67.

Pastor-Satorras, R. and Vespignani, A. (2001). Epidemic spreading in scale-free networks. *Physical Review Letters*, 86 (14): 3200–3203.

Paulus, P. B. and Nijstad, B. A. (2003). *Group creativity*. Oxford University Press, UK.

Phelps, C., Heidl, R., and Wadhwa, A. (2012). Knowledge, networks, and knowledge networks: a review and research agenda. *Journal of Management*, 38 (4): 1115–1166.

Risau-Gusman, S. and Zanette, D. H. (2009). Contact switching as a control strategy for epidemic outbreaks. *Journal of Theoretical Biology*, 257 (1): 52–60.

Rocha, L. E. C., Liljeros, F., and Holme, P. (2010). Information dynamics shape the sexual networks of iIternet-mediated prostitution. *Proceedings of the National Academy of Sciences*, 107 (13): 5706–5711.

Romero, D. M., Meeder, B., and Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, 695–704, Hyderabad, India. ACM Press.

Rosenberg, N. (1976). *Perspectives on technology*. Cambridge University Press, UK.

Ross, R. (1915). Some a priori pathometric equations. *British Medical Journal*, 1 (2830): 546–547.

Sampat, B. N. and Ziedonis, A. A. (2005). Patent citations and the economic value of patents. In Moed, H. F., Glanzel, W., and Schmoch, U., editors, *Handbook of Quantitative Science and Technology Research*, 277–298. Springer Netherlands, Dordrecht, Netherlands.

Schmookler, J. (1966). *Invention and economic growth*. Harvard University Press, Cambridge, MA, USA.

Schumpeter, J. A. (1934). *The theory of economic development: an inquiry into profits, capital, credit, interest, and business cycle*. Harvard University Press, Cambridge, MA, USA.

Signorile, V. and O'shea, R. M. (1964). A test of significance for the homophily index. *American Journal of Sociology*, 70: 467–470.

Silverberg, G. and Verspagen, B. (2005). A percolation model of innovation in complex technology spaces. *Journal of Economic Dynamics and Control*, 29 (1-2): 225–244.

Singh, J. and Fleming, L. (2010). Lone inventors as sources of breakthroughs: myth or reality? *Management Science*, 56 (1): 41–56.

Sobrero, M. and Roberts, E. B. (2001). The trade-off between efficiency and learning in interorganizational relationships for product development. *Management Science*, 47 (4): 493–511.

Solomon, S., Weisbuch, G., de Arcangelis, L., Jan, N., and Stauffer, D. (2000). Social percolation models. *Physica A: Statistical Mechanics and its Applications*, 277 (1-2): 239–247.

Spence, A. M. (1984). Industrial organization and competitive advantage in multinational industries. *The American Economic Review*, 74 (2): 356–360.

Stauffer, D. and Aharony, A. (1994). *Introduction to percolation theory*. Routledge, London, UK, 2d edition.

Stent, G. (1972). Prematurity and uniqueness in scientific discovery. *Scientific American*, 227 (6): 84–93.

Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., and Pietronero, L. (2012). A new metrics for countries' fitness and products' complexity. *Scientific Reports*, 2.

Tacchella, A., Cristelli, M., Caldarelli, G., Gabrielli, A., and Pietronero, L. (2013). Economic complexity: conceptual grounding of a new metrics for global competitiveness. *Journal of Economic Dynamics and Control*, 37 (8): 1683–1691.

Tedeschi, G., Vitali, S., and Gallegati, M. (2014). The dynamic of innovation networks: a switching model on technological change. *Journal of Evolutionary Economics*, 24 (4): 817–834.

Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. *The RAND Journal of Economics*, 21 (1): 172.

Ugander, J., Backstrom, L., Marlow, C., and Kleinberg, J. (2012). Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 109 (16): 5962–5966.

Valentin, F. and Jensen, R. L. (2002). Reaping the fruits of science: comparing exploitations of a scientific breakthrough in European innovation systems. *Economic Systems Research*, 14 (4): 363–388.

van den Bergh, J. C. (2008). Optimal diversity: increasing returns versus recombinant innovation. *Journal of Economic Behavior & Organization*, 68 (3-4): 565–580.

van Raan, A. F. J. (2004). Sleeping Beauties in science. *Scientometrics*, 59 (3): 467–472.

Vega-Jurado, J., Gutierrez-Gracia, A., and Fernandez-de Lucio, I. (2009). Does external knowledge sourcing matter for innovation? Evidence from the Spanish manufacturing industry. *Industrial and Corporate Change*, 18 (4): 637–670.

Vitali, S., Tedeschi, G., and Gallegati, M. (2013). The impact of classes of innovators on technology, financial fragility, and economic growth. *Industrial and Corporate Change*, 22 (4): 1069–1091.

Von Neumann, J. and Burks, A. W. (1966). *Theory of self-reproducing automata*. University of Illinois Press, Champaign, IL, USA.

Wagner, C. S. and Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy*, 34 (10): 1608–1618.

Walsh, V. (1984). Invention and innovation in the chemical industry: demand-pull or discovery-push? *Research Policy*, 13 (4): 211–234.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393 (6684): 440–442.

Werker, C. and Brenner, T. (2004). Empirical calibration of simulation models. Technical Report 0410.

Westbrock, B. (2010). Natural concentration in industrial research collaboration. *The RAND Journal of Economics*, 41 (2): 351–371.

Woolcock, M. (1998). Social capital and economic development: toward a theoretical synthesis and policy framework. *Theory and Society*, 27 (2): 151–208.

Wu, S., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Who says what to whom on twitter. In *Proceedings of the 20th international conference on World wide we*, 705–714, Hyderabad, India. ACM Press.

Zanette, D. H. and Risau-Gusman, S. (2008). Infection spreading in a population with evolving contacts. *Journal of Biological Physics*, 34 (1-2): 135–148.

Zeppini, P. and Frenken, K. (2015). Networks, percolation, and demand. Bath Economic Research Papers 38/15, Department of Economics, University of Bath, UK.

Zeppini, P., Frenken, K., and Kupers, R. (2014). Thresholds models of technological transitions. *Environmental Innovation and Societal Transitions*, 11: 54–70.

Zheng, M., Lu, L., and Zhao, M. (2013). Spreading in online social networks: the role of social reinforcement. *Physical Review E*, 88 (1): 012818.

Nederlandse samenvatting

Dit proefschrift omvat vijf studies over de sociale aspecten van kennis, en in het bijzonder de creatie en diffusie daarvan. Over het algemeen zijn creatie en diffusie van kennis inherent aan interacties in de maatschappij. In die zin zijn de resultaten van deze studies dus relevant en van toepassing in diverse sociale domeinen. De belangrijkste focus in dit werk ligt niet echter op specifieke kennis, maar meer op de onderliggende ideeën. Een terugkomend voorbeeld in de hoofdstukken van dit proefschrift is de dynamiek van wetenschappelijke ideeën: hoe ze worden beschouwd, geaccepteerd en verspreid binnen wetenschappelijke kringen.

De hoofdstukken verschaffen inzicht in de volgende volgordelijke elementen van kennisontwikkeling en -diffusie. Om te beginnen werken onderzoekers samen om wetenschappelijke kennis te genereren (hoofdstuk 2). Deze kennis wordt vervolgens verspreid binnen de wetenschappelijke gemeenschap (hoofdstuk 3). Onderzoekers die werken aan nieuwe ideeën en onorthodoxe methoden streven

er vaak naar gevestigde theorieën omverwerpen (hoofdstuk 4). Als zij daarin slagen, creëren zij doorbraken in wetenschap, hetgeen in sommige gevallen pas na verloop van tijd erkend wordt. Die gevallen noemen we 'Doornroosjes' (hoofdstuk 5). Ten slotte verschaft hoofdstuk 6 empirisch inzicht in de arbeidsverdeling in economische sectoren, in tegenstelling tot die in de productie van wetenschappelijke kennis. Deze verdeling is in kaart gebracht aan de hand van verschillende beroepsgroepen. Hoewel de context hier meer economisch van aard is, betreft deze laatste studie wel degelijk ook de interactieve manier waarop kennis ontstaat (vergelijk met hoofdstuk 2). Alle studies wijzen uit dat netwerk-theoretische concepten met succes toegepast kunnen worden in zowel empirische en theoretische onderzoeken naar creatie en diffusie van kennis.

We vatten hierbij de belangrijkste vindingen van de verschillende hoofstukken samen. Hoofstuk 2 presenteert een simulatiemodel van kenniscreatie in netwerken. Elkaar beïnvloedende actoren creëren kennis waarbij er sprake is van dubbele feedbackmechanismes: hun samenwerking is van invloed op de hoeveelheid kennis die ze creëren, wat vervolgens weer van invloed is op de structuur van hun samenwerkingen. De padafhankelijke ontwikkeling van de simulatieresultaten wijst op drie soorten patronen van co-evolutie tussen enerzijds de uitkomst van kenniscreatie, en anderzijds het aantal samenwerkende actoren. Deze relatie kan positief zijn (hoe meer betrokkenen, des te meer kenniscreatie), negatief (te veel betrokkenen werkt schadelijk) of onafhankelijk. De feedbackmechanismen tussen het samenwerkingsproces en het proces van kenniscreatie leiden tot enkele onverwachte resultaten. Causale verbindingen tussen samenwerking en de uitkomst van kenniscreatie vervagen. Een over het geheel genomen positief co-evolutie scenario kan gedreven worden door de aantrekkelijkheid van goed presterende actoren, zelfs als de betrokkenen niet uitzonderlijk productief zijn in termen van kenniscreatie. Het algemene doel van dit hoofdstuk was om de

aandacht te vestigen op deze feedbackmechanismes tussen endogene netwerkformatie en interactieve kenniscreatie.

Wanneer kennis eenmaal geproduceerd is, vind verspreiding plaats door middel van een netwerk. Hoofstuk 3 introduceert een percolatiemodel voor het analyseren van het effect van kennisverspreiding door middel van mond-tot-mond communicatie. De studie beschouwt kennisverspreiding in verschillende netwerkstructuren en populatietypes. Van specifiek belang is de invloed van sterke netwerkconnecties op de mate van kennisdiffusie, omdat deze invloed in de bestaande literatuur nog niet duidelijk is. Het belangrijkste resultaat is dat de zwakkere netwerkconnecties beter werken voor de diffusie van simpele en vanzelfsprekende ideeën, terwijl sterkere netwerkconnecties de diffusie van complexe en controversiële ideeën faciliteren in gemeenschappen die overeenkomstig gedachtegoed hebben.

De verspreiding van ideeën in een netwerk kan leiden tot een verschuiving in dominerende paradigma's. Hoofdstuk 4 beschrijft wetenschappelijke transities die ontstaan wanneer verschillende alternatieve theorieën strijden om dominantie. Elke nieuwe theorie maakt gebruik van 'defecten' in het originele paradigma. Deze kritieken tezamen leiden tot een toenemende steun ontstaat voor alternatieve theorieën. De nieuwe theorie die als eerste het punt van kritische fragmentatie bereikt (mogelijk omdat het aantal aanhangers van oorspronkelijke paradigma steeds sterker versplinterd) neemt vervolgens alle voorstanders van alternatieve theorieën op en wordt zelf dominant, en vervangt daarmee het oorspronkelijke paradigma.

Een mogelijke aanleiding voor een paradigmaverschuiving is de ontwikkeling van een technologische doorbraak. Hoofdstuk 5 verkent de vertraagde erkenning van kennisdoorbraken door middel van de identificatie en empirische analyse van 'Doornroosjes' in technologie. Deze uitbreiding van het concept van Doornroosjes van het domein van de wetenschap naar dat van de technologie is een

nieuwe aanpak, en laat zien dat niet alleen wetenschappelijke maar ook radicale technologische ontwikkelingen er soms niet in slagen om zich direct te verspreiden. Doornroosjes zijn vaak ontwikkelingen die plaatsvinden buiten de heersende technologische trajecten. Dit blijkt doordat hun octrooien, in vergelijking met andere octrooien met veel citaties, een relatief klein aantal referenties naar bestaande technologie hebben, en tevens relatief veel auteurs kennen. Doornroosjes kunnen echter ook op gefaalde diffusie wijzen, wat gesuggereerd wordt door de geringe ervaring van hun auteurs, en het feit dat verlate doorbraken in relatief vergaand gecodificeerde technologische domeinen onwaarschijnlijker zijn.

Ten slotte bestudeert hoofdstuk 6 de kennisbasis van economische sectoren door de netwerkstructuur van hun functionele arbeidsverdeling in kaart te brengen. De studie maakt gebruik van recente netwerkanalyse-technieken om nieuwe indicatoren te genereren voor de functieverscheidenheid in sectoren en de alomtegenwoordigheid van die arbeidsfuncties. De negatieve relatie tussen deze verscheidenheid en alomtegenwoordigheid laat zien dat de variatie in functiegebruik op een specifieke manier verweven is met de intenstiteit daarvan. Anders gezegd, juist die sectoren welke een veelheid aan functies herbergen kennen de meest unieke functies, terwijl de meeste industrieën juist de voornamelijk algemenere functies gebruiken. Verscheidenheid en alomtegenwoordigheid zijn ook voorspellers voor de productiviteit van sectoren. Zeer productieve sectoren gebruiken veel functies. Verder zijn de functies die de meeste productiviteit brengen de algemenere functies; de functies die minder vaak extern vervuld worden. Een voor de hand liggend vervolg op deze studie is om de analyse te herhalen in de context van wetenschap, door te kijken naar de verscheidenheid aan functies in verschillende wetenschappelijke disciplines.

Acknowledgements

This thesis has been made possible thanks to the patience and collaboration of very many people. I would like to specially thank my PhD committee, for reading my manuscript and collaborating in its improvement. Conversations with my supervisors have been crucial for my development as a researcher and for my personal growth. Ximo first encouraged me to follow any research path I would enjoy, and he got a little disappointed that I liked knowledge networks much better than our less conventional options (which included the Marvel universe and amateur marathon running). Koen arrived to my PhD project right at the point where I had to change my proposal (the original one was so good it had already been published) and encouraged me to attend to as many conferences as possible, allowing me to build an international profile and many rewarding collaborations. With Paolo I could communicate in my own mathematical language, and he taught me how to translate it to appeal to different audiences (I am still learning, but I have gotten better thanks to him).

The people at the institutions where I have been during these years have made my experience as a PhD student a greatly enjoyable process. I am most grateful

to all my colleagues at Ingenio for their encouragement and their help, specially to Pablo, Antonio, Isabel, Ester, Carlos, François, Mabel, Óscar, Alfredo, Ana and Anabel. Davide has been a most kind and systematic coauthor, I have learnt so much from him. Richard and Julia made my arrival easy and fun, while Inma, Sergio and Anna have made my last and exhausting months there happy and very amusing.

It is the merit of the colleagues at Eindhoven and Utrecht that I have enjoyed so much my stays there. Yuti, Boukje, Julia and Katerine are some of the best roommates that anyone can hope for. Alco and Denise actually made me want to go to the office when I was in Utrecht and are now taking very good care of my plant. Matthijs (to whom I still owe a paella) and Rudi (who introduced me to TWM) made some of the nicest memories I have from the Netherlands. I have been very lucky to spend time with Daniel and Roman, whether dancing bachata or working on random heavy metal projects. I have had a huge lot of fun with you all.

A lot of credit is due to the affectionate, enormous support of my people at home. Papá, Mamá, you introduced me to research and assisted me throughout each and every step that I have ever taken. Alicia y Norat, you are the best boardgamers ever and you have to come back for a trip. Carmelo i Sesi, heu sigut com uns segons pares per a mi, vos estic molt agraïda per els tapers i les fotos i el reiki i tot el vostre suport. Isa, Carme, Irene, María, Thais: menos mal que he tenido las sesiones de cafetería. Qué hubiese sido de mí.

Finally, this thesis is as much mine as it is my husband's. You should be getting a PhD when I defend, Abel. During the last months you struggled and suffered and always, always managed to make everything so easy. I am so very grateful. There is absolutely nothing I could ever say that will even come close to how grateful I am for having you. Te uic.

Curriculum Vitae

Elena M. Tur was born the 11th August 1987 in Valencia, Spain. After starting a degree in Mathematics at the University of Valencia, she enjoyed a Introduction to Research grant of the Spanish Research Council that inspired her to follow an academic career. She then decided to enroll in a Statistics degree and later in a Master in Industrial Economics, both awarded with the highest grade of distinction. Her performance during her Statistics studies was recognized with an accessit to the 2011 National Best Graduate Award.

During her Master studies, Elena was awarded a PhD scholarship of the Spanish Research Council. She then joined the research team at Ingenio (CSIC-UPV), where she had already been working during a summer studentship. She has been a visiting PhD student at the Eindhoven University of Technology and the University of Utrecht. After finishing her PhD thesis, Elena received a Broman scholarship from the University of Gothenburg. She is currently an assistant professor at the Eindhoven University of Technology.