DISS. ETH Nr. 25960

The structure, exchange, and transfer of knowledge in socio-technical systems

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

presented by
GIACOMO VACCARIO

M. Sc. in Physics, Sapienza, Università di Roma
MASt. in Mathematics, University of Cambridge

born on February 26, 1990

citizen of Italy

accepted on the recommendation of
PROF. DR. DR. Frank Schweitzer, examiner
PROF. DR. Andreas Pyka, co-examiner

2019

ETH zürich

Acknowledgements

The first page for what it is written at the end: the acknowledgements.

I thank Professor Frank Schweitzer for his supervision, made of clear suggestions and critical observations, that offered an indispensable help to continue in the right direction. When comparing the person I am today with the one I was four years ago, I see that the doctorate experience completed with him has been essential. He allowed me to grow at the professional level by improving my critical thinking and ability to analyze problems rigorously. Moreover, his supervision has been precious for my personal growth. He taught me not only how to recognize my strengths and weakness, but also to respectively use and overcome them. Thanks, Frank, for all this.

Thanks to my close collaborators: Luca Verginer, Mario V. Tomasello, Manuel S. Mariani, Matùš Medo, Nicolas Wider, and Claudio J. Tessone. Without their collaboration, the research performed and discussed in the present thesis would not be here.

Also, I am thankful to Professor Andreas Pyka, whose work inspired part of my research, and who has agreed to be a co-examiner of this thesis.

Thanks to all the colleagues at the Chair of Systems Design: Giona, Vahan, Simon, Ingo, Yan, Rebekka, Antonios, David, Emre, Adiya, Pavlin, Ambra, Christian, Christoph, Ramona, and Laurence. From our daily discussions and their experience, I obtained many inputs and suggestions. All these have contributed to form a complete and coherent vision of my research. In particular, a big thanks goes to Giona and Vahan for helping with mathematical and technical details, and to Simon, for his invaluable remarks on how to address scientific problems from a different perspective. Also, I have to thank Ingo as his help has been crucial in the starting phase of this thesis. He motivated me to pursue new research paths and pushed me to improve myself continuously.

Thanks to Andrea, Giovanni, Paolo, Federico, and Rushd, old colleagues from Rome and new friends in Zurich. By always feeding me new food for thoughts, they made me surpass my past working and studying experiences. Moreover, with intensive sport sessions, and generous dinners, that usually went with

a good bottle of wine, they have enriched my stay in Zurich not only with moments of recreation but also, and more importantly, with great comfort.

Thanks to my parents for their continuous presence and support during this long adventure. Talking and discussing with them, it gave more transparent motivations and more comprehensive breadth to my work.

I thank Tommaso, my brother, who supported me when I felt more insecure and helped me to find self-confidence.

Last, but not least, thanks to Valentina, my wonderful girlfriend, who always assisted and advised me. From rethinking an image to correcting a paragraph, or in reviewing research ideas during our walks along the Limmat. Thanks for staying next to me and for continuously providing your support. The last thanks to you that are the first in my thoughts.

Contents

Abstract

This thesis aims to improve our understanding of the role of knowledge in economics and science. We analyze collaboration activities in these two domains, and show how the interactions among firms and among scientists influence the structure and the exchange of knowledge. We also model how the knowledge of these actors defines their collaboration activity. We show that knowledge is not only a consequence, but also a determinant of the collaborations. To capture this interplay, we combine a statistical analysis of patent and publication data with agent-based models of collaboration activities.

We follow a data-driven approach to study the structure, exchange, and transfer of knowledge. Specifically, using publication data we proxy the structure of scientific knowledge by reconstructing the citation network between publications. On this network, we quantitatively show that citation patterns strongly differ across time and scientific fields. We also identify the different knowledge of scientists, and quantify their knowledge exchange occurring during collaborations. Similarly, we use patent data to identify firms' knowledge and the knowledge exchange between firms involved in R&D alliances. Then, to study the transfer of knowledge, we re-construct scientists' career paths by tracing their affiliations reported on their publications. With these paths, we construct the global migration network of scientists at city level, and analyze its topological properties.

After analyzing collaborations activities, the exchange, and the transfer of knowledge, we reproduce these using agent-based models that we calibrate an validate against real-world data. In order to capture the very different processes behind these phenomena, we develop three different models. Precisely, to model collaborations activities among firms and their subsequent knowledge exchange, we combine and extend two existing models that captured only one of these phenomena each. Our a new model, instead, is able to simultaneously reproduce both these phenomena. To show how the knowledge differences between scientists determine their collaboration activities, we develop a second model that takes as input only these differences. Then, to model the transfer of knowledge, we develop a third agent-based model that reproduces scientists'

migration at city level and the observed topological properties of the global migration network.

Finally, we show that citation patterns between journals and scientists' career paths are better modeled by a new mathematical framework defined by higher-order networks than by traditional network models. By this, we challenge the application of the traditional network perspective to model the flow of knowledge between journals and the transfer of knowledge across research institutes.

Sintesi

La seguente tesi intede mettere a fuoco come si esplica la funzione della conoscenza, nell'ambito dell'econonomia e della scienza. Innanzi tutto analizziamo le attività di collaborazione in questi due domini e mostriamo come le interazioni tra le imprese e tra gli scienziati influenzano la struttura della conoscenza e il suo scambio. Modelliamo anche come la conoscenza di questi attori definisce le loro attività di collaborazione. Si dimostra quindi che la conoscenza non è solo una conseguenza, ma anche una determinante delle collaborazioni. Per cogliere questa reciproca influenza, combiniamo l'analisi statistica di dati di brevetti e pubblicazioni scientifiche con modelli ad agenti per modellare le attività di collaborazione.

Per esaminare la struttura, lo scambio e il trasferimento delle conoscenza, seguiamo un approccio data-driven, ossia basato sui dati. Nello specifico, utilizzando dati riguardanti pubblicazioni scientifiche, individuiamo la struttura della conoscenza scientifica ricostruendo la rete di citazioni dalle pubblicazioni. Su questa rete dimostriamo quantitativamente che le citazioni ricevute da pubblicazioni differiscono fortemente in base all' età delle pubblicazioni e dal loro ambito scientifico. Identifichiamo anche le diverse conoscenze degli scienziati e quantifichiamo il loro scambio di conoscenza dovuto alle loro collaborazioni. In modo analogo, utilizziamo i dati sui brevetti per identificare le conoscenze delle imprese e lo scambio di conoscenza tra imprese coinvolte in alleanze di R&D. Dopoddiche' ricostruiamo le carriera degli scienziati tracciando le loro affiliazioni riportate sulle loro pubblicazioni, in modo da studiare il trasferimento della conoscenza. Con queste carriere, ricostruiamo anche la rete di migrazione degli scienziati a livello di città e ne analizziamo le proprietà topologiche.

Dopo aver analizzato le attività di collaborazione, lo scambio e il trasferimento di conoscenza, riproduciamo questi fenomeni utilizzando modelli ad agenti che calibriamo e validiamo con dati provenienti dal mondo reale. Al fine di catturare processi molto diversi alla base di questi fenomeni, sviluppiamo tre diversi modelli. Per modellare le attività di collaborazione tra le imprese e il loro successivo scambio di conoscenza, uniamo ed estendiamo due modelli esistenti, ciascuno capace di modellare solo uno di questi due fenomeni

separatamente, mentre il nostro nuovo modello è in grado di riprodurre simultaneamente entrambi i fenomeni. Calibriamo e convalidiamo questo modello confrontandolo con i dati del mondo reale e mostriamo come esso cattura bene le proprietà macroscopiche osservate in questi fenomeni. Per mostrare come le diverse conoscenze degli scienziati determinano le loro attività di collaborazione, sviluppiamo un secondo modello che prende come input solo le loro differenze. Mentre, per modellare il trasferimento della conoscenza, sviluppiamo un terzo modello ad agenti che riproduce la migrazione degli scienziati da città in citta' e le proprietà topologiche della rete di migrazione osservata.

In fine, mostriamo che per rappresentare le citazioni tra riviste scientifiche e le carriere degli scienziati e' meglio utilizzare dei nuovi modelli matematici basati sulle *higher-order networks* rispetto ai tradizionali modelli di rete. Con ciò mettiamo in dubbio l'applicazione dei tradizionali modelli di rete per modellare il flusso di conoscenza tra riviste scientifiche e il trasferimento di conoscenza tra istituti di ricerca.

Chapter 1

Introduction

Knowledge is considered one of the key production factors determining wealth in modern capitalist societies. The competitive advantage of firms depends more and more on their investments in knowledge-based capital, while a large share of the market value of many leading firms reflects their knowledge assets, such as patents and know-how. However, the importance of knowledge is not solely related to economics. Science, too, creates new knowledge making way for scientific progress, which has a direct impact on many aspects of our society. As we live in a digitally-dominated era, thanks to the digitalization, we collectively store and process, and we retrieve an exhaustive amount of information. For the first time in human history, we easily access exabytes of information coming from news, books, scientific publications, patents, product descriptions, online forums, and many other sources. By navigating and processing this information, we extract new knowledge that we use not only to create new products and services, but also to make discoveries about the functioning of our society.

The above statements clearly show that knowledge places a central role in economics and science. Note that in these two domains, knowledge is created more and more in a collaborative manner. This collaboration effort is evident from the increasing number of research and development alliances established between firms, or the increasing number of discoveries obtained by large teams of scientists. These collaborations are not bound by geography, rather they connect firms, or scientists, across nations and continents. Moreover, they

are increasingly connecting firms working in different economic sectors, and scientists of varying expertise.

The increasing number of collaborations devoted to the creation of knowledge challenges the traditional approach to studying knowledge as a property owned by single individuals. In other words, given that knowledge is created as a joint effort of many scientists or even of groups of firms, why should we study knowledge just by looking at individuals separately? In this thesis, we argue on the necessity to study knowledge as dependent on the *collection of interactions* among knowledge actors, such as scientists and firms.

In order to study the collection of interactions determining knowledge, we take the complex system perspective to analyze the elements constituting knowledge and their interactions – all at the same time. The analyzed elements are knowledge artifacts, like patents and publications, and knowledge actors, such as firms and scientists. The interactions between these elements range from citation relations between knowledge artifacts to collaboration activities among knowledge actors.

To analyze knowledge artifacts and their relations, we build on methods coming from the scientific fields of information science and scientometrics. Scholars from these fields have developed various criteria and indicators to classify, rank, and retrieve patents, scientific papers, and journals. Developing new criteria and indicators to perform these tasks is the first objective of this thesis.

As the second objective of this thesis, we study knowledge actors and how they collectively produce knowledge. We do this by developing agent-based models, i.e., using computational models that simulate the interactions of autonomous agents to reproduce collective behavior. Note that our agent-based models will not be just computer simulations. By theoretically grounding the microscopic rules of our models and by calibrating and validating them against real data, these models will not only reproduce observed collaborations patterns but also allow for understanding.

In the following, we present the working definition of knowledge that we adopt and define in which systems we investigate knowledge (see Sect. 1.1). Then, we present our research agenda by describing the research gap that we have identified during the preliminary phase of the thesis (see Sect. 1.2). Additionally,

in Sect. 1.3 we provide seven research questions, answers of which contribute to filling the identified gap. In Sect. 1.4, we describe in which thesis chapters we address the different research questions.

1.1 What is knowledge?

We aim to investigate knowledge in the domains of science and economics. Though an unambiguous definition of knowledge is disputed, we now discuss the meaning of this concept in both domains.

In economic theory, knowledge would typically be considered as "human capital" [18, 102, 137, 183, 207, 246]. This "capital" should not only be seen as a bare collection of goods, for it is often embedded in a social context that can strongly affect its properties and value [39, 51, 197]. To include this characteristic, some scholars consider knowledge as *"a dynamic framework connected to cognitive structures from which information can be stored, processed and understood"* [96, 197]. By this, we explicitly state that knowledge involves cognitive structures, such as humans, that can be embedded in a social context and hence be affected from it.

The above concept of knowledge comes from economics [96], and it is also accepted by scholars modeling the structure and evolution of science [197]. For this reason, it well matches the general aim of this project, which is investigating knowledge in the domain of science and economics.

Data, information and knowledge When defining *knowledge* we used the term *information*. These are two distinct but connected concepts. A detailed discussion about their differences goes beyond the scope of this thesis. At the same time, let us clarify how we interpret and use these terms and how they are related to *data*. *Data* is a collection of values and does not have a specific meaning by itself. For example, a list of 100 dates alone is just some data. *Information* is a collection of values whose meaning depends on a question asked. For example, if we ask "Which are the publishing dates of Einstein's articles?" and receive a list of 100 dates as an answer, this list is not just data, but information. In other words, information is an answer to a question, while

data is not. Lastly, by storing, processing, and understanding information, we can discover and acquire *knowledge*. For example, by investigating the age at which Einstein published his most successful articles, we could gain some knowledge about his scientific productivity.

Academia and the R&D alliances as socio-technical systems. In modern society, there are many systems in which social and technical aspects are strongly intertwined. Prominent examples are academia, meaning scientists with their research activities, and R&D alliances, meaning partnerships arranged by firms to promote their R&D activities. In these systems, the social actors (scientists and firms) influence each other depending on both their social and technical characteristics. For example, in academia, scientists' technical skills and knowledge are essential for their scientific impact. At the same time, scientists' scientific impact is correlated with their centrality in the social network reconstructed using their co-authorship activities [194]. In economics, firms select R&D partners depending on their social and technological capitals [32, 82, 165, 185]. This interdependence between social and technical aspects makes academia and the R&D alliances two socio-technical systems.

In this thesis, we will analyze the structure, exchange, and transfer of knowledge in socio-technical systems.

1.1.1 The structure, exchange, and transfer of knowledge.

There exists a vast literature studying the structure, exchange, and transfer of knowledge. Important contributions are [7, 91, 101, 102, 109, 151, 152, 157, 225] in the economic domain and Boyack *et al.* [30], De Domenico *et al.* [48], Gargiulo *et al.* [69], Lotka [123], Noyons and van Raan [155], Price [170, 171] in the scientific domain.

In particular, Jaffe and Trajtenberg [102], Tomasello *et al.* [225] use patent data to quantify innovation and measure knowledge flows. Indeed, knowledge can be encoded in patents, and by this, patents become knowledge artifacts. In the attempt of establishing different types of knowledge, Polanyi

[166] distinguished between *explicit* and *tacit* knowledge. By following this point of view, we assume that explicit knowledge is encoded in knowledge artifacts [96, 151, 152, 166]. Hence, we can investigate the structure and evolution of knowledge using patents and their relations. Note that the possible relations between patents are many. Obvious examples of relations include the direct citations between patents, while less obvious ones are the geographical distances between the inventors filing the patents.

In the field of *scientometrics* and *information science*, scholars use an equivalent approach to study explicit knowledge in science. They consider scientific publications as knowledge artifacts. By analyzing these artifacts and their relations, they drawn map of scientific knowledge [30, 120, 210] and model its evolution [26, 48, 170]. Additionally, scientific publications are traditionally grouped in journals and hence, journals create a mesostructure organizing knowledge artifacts. Developing indicators for ranking and classifying journals is an ongoing effort also because these indicators are official evaluation tools in various countries.

In both the economic and scientific domains, knowledge is created more and more in a collaborative manner. In the economic domain, we have withstood a shift towards open innovation [37] and an increasing number of R&D alliances [5, 85]. To interpret this, we follow the knowledge-base view of firms and considers firms' knowledge a fundamental resource of their competitive advantage [56, 107, 253]. In order to expand or change their knowledge base, firms establish alliances among each other allowing for knowledge to be exchanged [46]. Then, modeling these alliances together with the possible knowledge exchange, it is central to understand innovation processes in our economies.

Similarly, in the scientific domain, scientists also collaborate in order to produce scientific progress and publish their results in journals. Indeed, the most extensive available data about the scientific collaborations comes from bibliographic databases containing co-authored scientific publications. From the analysis of these, there is strong evidence that science is done more in more in teams [256]. This result is partially explained by the fact that by collaborating individuals can stimulate each other, share complementary resources and pro-

duce original works [230]. In other words, collaborations allow for knowledge to be exchanged, and this can increase the quality of the work done. To a limited extent, this is verified by the fact that scientific publications co-authored by larger teams also tend to receive higher attention [113]. At the same time, there are also pieces of evidence challenging the above claims. For example, Wu *et al.* [255] show that smaller teams produce scientific works that are more disruptive and Wagner *et al.* [243] failed to show that more international collaborations produce more original articles. Moreover, it is still an open problem to quantify how much knowledge is exchanged thanks to collaborations.

Even though knowledge is often treated as an abstract quantity, it is still coupled to the real, physical space. Recall that knowledge is connected to cognitive structure, such as individual human beings, and individuals do not have access to all the possible information [211]. Depending on their locations, individuals have access to different local information that they can process. From this, they obtain a "knowledge of the particular circumstances of time and place" [91]. Individuals then own this knowledge and they can transfer it thanks to their physical movements. Note that with their movements, individuals transfer both their explicit and tacit knowledge. For this reason, the physical movement of humans is an essential form of knowledge diffusion studied in various works [69, 115, 241].

In this thesis, we will focus on the mobility of scientists to proxy the *transfer of knowledge*. We are tempted to use expressions such as *mobility of knowledge* or *migration of knowledge*. At the same time, these terms could be easily misinterpreted. The term "migration" usually refers to the movement of humans to a new country or the seasonal movement of animals. Whereas the term "mobility" is mostly used in the field of physics to describe how quickly particles move in materials (e.g., electron mobility). In a similar sense, the term "mobility" is also used to discuss human dynamics (e.g., urban mobility). To avoid such confusion, we define the *transfer of knowledge* as the spreading of knowledge occurring thanks to the physical movement of scientists or inventors. For example, scientists move and transfer knowledge from one university to another, or inventors move and transfer knowledge from one firm to another.

Our analysis of the structure, exchange, and transfer of knowledge follows a research agenda based on the research gap presented in the next section.

1.2 Research gap

To define the research gap, we start by introducing two broad classes of problems linked to two overarching questions. The first class of problems is related to the question of *how knowledge artifacts are linked to each other, ranked and filtered in repositories*. Examples of these repositories are patent and scientific publication databases. The second class of problems relates to the question of *how knowledge is exchanged and transferred*. Indeed, knowledge is not only produced and encoded in knowledge artifacts, like patents or scientific publications, but it is also exchanged by human beings, and it diffuses in our society. R&D alliances among firms and co-authorship of papers among scientists are examples of activities favoring knowledge exchange. While the physical movement of inventors or scientists is an example of knowledge transfer.

Answering the questions highlighted in the above paragraph is a multidisciplinary effort made by many researchers. In particular economists, bibliometricians and computer scientists contributed by proposing tools for managing knowledge artifacts, modeling knowledge exchange and its transfer [1]. Successful tools have been developed by applying network-analytic methods [31, 38, 69, 164, 225]. The use of such methods was possible as the data could be represented using a network abstraction. For example, we can represent a scientific publication database as a *journal-journal citation network*, that is a network where a node is a scientific journal and a link between two journals is a citation between two papers published in these journals [164]. Similarly, data about scientists' careers [2] can be represented as *faculty hiring network*, which is a directed network where each node is a university, and a directed link represents a scientist graduating from a (source) university and, becoming a faculty member in a target university [38]

[1] When writing knowledge transfer, we always refer to the scientists mobility, i.e., to scientists that move across firms, institution, cities or countries, unless otherwise specified.

[2] Here, we exactly mean the personal career trajectory of a scientist in his/her professional life.

Even though tools developed using network-analytic methods have been proven to be useful, recent advances in data mining and network theory raised concerns about their naive application to complex data [34, 129, 201, 262]. In the following sections, we will give an overview of three critical concerns.

1.2.1 Network-analytic methods and citation data

One established method to extract knowledge from patent and scientific publication data is *citation analysis*. It is used to investigate citation patterns in order to disclose the properties of documents. In citation analysis, we use network-analytic methods as we can represent documents and their citations as a *citation network*. In this network, nodes are documents, and links are citations among documents. When representing citation data as a network, we discard many proprieties of the documents, such as their age or their topic of pertinence.

At the same time, the cumulative number of citations received by documents strongly depends on their age and topic of pertinence [205, 240]. Then a question arises: *how can we use citation analysis to compare documents of different age and belonging to different topics?* Answering this question is important for many reasons. For example, we increasingly use citation analysis for research evaluation, but we cannot use it fairly to compare researches who publish on different topics.

Various works proposed an answer to the above question (see [247] for a recent review). A simple, yet successful method is to divide each document's citation count by the mean number of citations for documents on the same topic published in the same year [179]. By this procedure, we obtain a new indicator called *relative citation count*. Radicchi *et al.* [179] claim it can be used as "*an unbiased indicator for citation performance across disciplines and years*" for scientific publications. Subsequent works challenged this finding [6, 249], which leaves the debate on age- and field- normalization procedures still open. Besides, to the best of our knowledge, a statistical test to assess if indicators are simultaneously age- and field-normalized is still missing.

1.2.2 Collaboration networks and knowledge

Researchers from many disciplines have been addressing the long-lasting question of how scientists and firms collaborate to create knowledge. In the economic domain, they have investigated the mechanisms behind the formation of R&D alliances [169], the complex networks they generate [187, 224], and the way their evolution can be modeled [65, 110, 175]. In particular, Tomasello *et al.* [224] successfully use an agent-model to reconstruct the R&D network between firms and hence, their collaboration patterns. Building on such a network model, [225] investigates how firms exchange knowledge during collaborations. Interestingly, the authors find that firms exchange knowledge at a low rate, meaning that knowledge is rather a determinant than a consequence of collaborations. This result is in line with the fact that firms use collaborations to access new knowledge [153]; however, then firms do not use this knowledge to expand their knowledge base.

Tomasello *et al.* [225] assign a knowledge position, i.e., a vector, to firms by looking at their patent portfolios. This approach is similar to the one proposed in SKIN models where the knowledge of a firm is contained in a "kene" [75]. The main difference between SKIN models and the models of Tomasello *et al.* [225] is that in the former each agent has its knowledge space defined by the kene, while in the latter agents have different positions in a shared knowledge space. Additionally, in [75] agents' kenes are composed of three parts that represent firms' capability (i.e., technological or business domain), ability, and expertise. While in Tomasello *et al.* [225], in a data-driven spirit, the authors model only firms' expertise using their observed patent portfolios.

In particular, Tomasello *et al.* [225] consider the main eight sections of the International Patent Scheme (IPC) to classify firms' patents, and hence, the knowledge space has eight dimensions. Such choice raises two technical, but rather critical problems. *First*, the model of [225] is a particular type of a n-vector model in eight dimensions [218]. Then, its properties and behaviors are linked to its dimensionality. This means that the results obtained using the IPC might not hold when using another classification scheme with a different number of dimensions. *Second*, it is reasonable to doubt that the eight dimensions of the IPC sections can adequately describe the knowledge space: i) IPC

sections are broad (i.e., sections include diverse technologies that are unlikely to share the same knowledge base), ii) some technological fields are dispersed across several IPC sections (e.g., food-chemistry related patents may be found in two different sections). To establish if the result of [225] holds when using more refined patent classification schemes with a higher dimension is still an open question.

Also, the question of how the actors of other domains collaborate and exchange knowledge is unanswered. One could argue that the answer to such question changes with respect to the actors and the domain of activity, but there may also be evidence for common features across different domains. Indeed, in a recent article, we show that a unified modeling approach can reproduce and explain the structural and the dynamic features of collaborations in different domains [227]. Such work uses the same agent-based model of [224] to investigate the collaboration patterns in economics and science. Precisely, the authors focus on collaborations aimed at the production of new knowledge: R&D alliances among firms from 6 different industrial sectors and co-authorship activities among scientists from 6 distinct disciplines of *physics*. Additionally, the idea of using "kenes" to represent knowledge was successfully applied to both the academic [73] and economic domains [75].

Building on the result and the methods presented in [73, 75, 225], it is interesting to ask how much knowledge scientists exchange during co-authorship activities. In particular, it is interesting to check if the result of [225], namely that knowledge is rather a determinant than a consequence of collaborations, holds in the scientific domain.

1.2.3 The loss of temporal correlations

The third concern arises as we often need to make the *transitivity assumption* in order to use network-analytic methods. By transitivity assumption we mean the following: when inferring (from data) the existence of links from a to b and from b to c, we automatically permit the existence of a path from a to c via

b [3]. Recently, [204] have shown that this assumption is not justified in many real-world systems as non-trivial temporal correlations of events invalidate it. For this reason, network-analytic methods can be inadequate to investigate time sequences of events.

Recent advances in data mining and network theory overcame some of the limitations caused by the transitivity assumption. In particular, in [201, 203, 204], the authors show how to correctly construct networks and study diffusion processes by using data containing time sequences of events. Also, [201] provides a statistical test for doing model selection that accounts for possible temporal correlations (present in the data). And in this test, a network is one of the candidate models. Hence, we can now verify when it is justified to interpret data using a network abstraction and when it is not.

With the methods provided in [201, 203, 204], we can investigate two sets of problems: *old* problems that might have been tackled incorrectly by using a network perspective (i.e., relevant temporal correlations might have been discarded), and *new* problems for which temporal ordering is critical, but have not been tackled yet. In both these sets, we are interested in those problems related to knowledge and its transfer.

1.3 Research Questions

We now provide 7 research questions (RQs) whose answers will contribute to fill the above stated research gap.

1.3.1 The structure of knowledge

RQ1: Assessing multiple normalizations. We need to age- and field-normalize citation-based indicators in order to compare documents of different

[3] From a mathematical point of view this assumption is equivalent to the following statement: after inferring from the data a set of nodes and a set of links to construct a network and its adjacency matrix, the n^{th}-power of the adjacency matrix is the set of paths of length n allowed on the network.

ages and from different fields. How can we assess that these indicators have been *simultaneously* age- and field- normalized?

RQ2: Developing a new normalization procedure. As the procedure of [179] failed to correctly age- and field-normalize citation count, and to the best of our knowledge, there are no better ones, how can we develop a better one?

RQ3: Developing a new knowledge order. How can we include time-correlations present in citation data to develop citation-based indicators?

1.3.2 Collaborations and knowledge exchange

RQ4: Knowledge as a determinant of alliances among firms. The results of [225] indicate that knowledge is rather a determinant than a consequence of R&D collaborations. How does this result change when using different methods to quantify knowledge?

RQ5: Knowledge as a determinant of collaborations among scientists. How can we extend the model and the analysis of [225] to the scientific domain?

1.3.3 The transfer of knowledge

RQ6: Temporal correlations in the transfer of knowledge. How should we model scientists' career paths in order to retain temporal correlations and a network perspective?

RQ7: A new agent-based model for knowledge transfer. Let us assume that temporal correlations in scientists' career paths break the transitivity assumption. Then, how can we use an agent-based model to reproduce these types of paths?

1.4 Structure of the thesis

This thesis is divided into 10 Chapters. In this first introductory chapter (Chapter 1), we have defined the scope of the thesis and our multidisciplinary approach. To do it, we have introduced our working definition of knowledge, determined the system we are going to consider, and identified a research gap together with seven research questions. In the next chapter (Chapter 2), we introduce the data and methods that we will use to address the posed research questions. Then, we will address these questions in the subsequent chapters.

The first three research questions are related to the structure of knowledge and are addressed in Chapter 3 (**RQ1** and **RQ2**) and Chapter 4 (**RQ3**). In these two chapters, we proxy the structure of knowledge space by using large citation networks of scientific publications. By understanding how centrality measures and clustering methods respectively rank and group publications on citation networks, we learn about the structure of the knowledge space. Precisely in Chapter 3, we discuss the field and time normalization problem of scientometrics indicators based on citation data. In our discussion, we define i) a statistical method that allows quantifying biases of indicators and ii) a normalization procedure that suppresses the observed biases. In Chapter 4, we study the temporal sequences in which publications cite each other at the journal level. With our study, we identify statistically significant sequences of citations between journals, and we use these sequences to rank journals and better capture their similarity. Note that Chapter 3 is based on [232], while Chapter 4 on [234].

In Chapters 5, 6, and 7, we address two research questions about the role of knowledge in determining collaborations in socio-technical systems (**RQ4** and **RQ5**). In Chapter 5, we embed firms in a knowledge space depending on their patent portfolios and model the formation of alliances among firms and their consequent knowledge exchange. Our analysis follows the approach of [224, 225, 226]. Note that differently from these works i) we consider two different patent classification schemes as firms' embedding and ii) when modeling firms' knowledge dynamics, we model *only* the knowledge exchange. By this, we can reliably study how much knowledge is exchanged during R&D alliances and

hence, how much firms move in the knowledge space. In other words, we investigate whether firms' knowledge should be regarded as a consequence or a determinant of collaborations (**RQ4**). The work presented in Chapter 5 is based on [233]

After analyzing R&D alliances, we turn our attention to collaborations among scientists. To do this, we first perform an extensive analysis of the interplay between knowledge and scientific co-authorship activities in Chapter 6. In this analysis, we also propose a new measure to capture the empirical knowledge exchange between scientists. Then, in Chapter 7, we present an agent-based model where agents represent scientists and choose their collaborators only depending on their knowledge. By this, we study the role of knowledge in determining scientists' collaborations (**RQ5**). Chapters 6 and 7 are unpublished.

Finally, in Chapter 8 and 9 we address **RQ6** and **RQ7**, i.e., we analyze and model scientists' career trajectories across the geographical world. Recall that, by focusing on scientists, we have the opportunity to capture both explicit and tacit knowledge. Hence, when we analyze scientists' career trajectories across universities, cities, or countries, we are also analyzing the transfer of knowledge across these locations. Additionally, in this chapter, we also reconstruct the scientists' mobility network at the city level and check whether there are statistical significant temporal correlations in scientists' career trajectories at the affiliation, city, and country level (**RQ6**). In Chapter 9, we model scientist career trajectories at the city level (**RQ7**). These last two chapters are based on [231].

We conclude our thesis with Chapter 10 where we present our conclusions. In these conclusions, we link the different chapters in order to have a new aggregated picture of the interplay between knowledge and socio-technical systems. After this last chapter, we also report various Appendices that complement the results reported in the thesis.

Chapter 2

Data and methods

Summary

In this chapter, we present the data and the modeling approach used for this thesis. Starting from the data, we describe seven large scale databases and explain how we will use them to study knowledge in socio-technical systems. After describing the data, we formally introduce the network perspective together with necessary network measures and algorithms. We will use this perspective to represent the systems analyzed in the following chapters. Finally, we describe and discuss the difference between four micro-based modeling approaches for networks: temporal exponential random graph model (TERGM), stochastic actor-oriented model (SAOM), relational event model (REM), and data-driven agent-based model (ABM). In our discussion, we emphasize the advantages and disadvantages of the different modeling approaches and motivate why we use data-driven agent-based models in this thesis. [1]

[1] Based on [227, 231, 232, 233, 234]

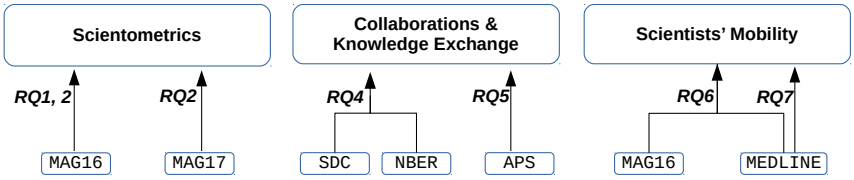


Figure 2.1: The relation between databases used and **RQ**s.

2.1 Data

We summarize the data used to address the research questions introduced in Chap. 1. We divide the data accordingly to the previously identified macro-search areas: *Scientometrics*, *Collaboration and Knowledge Exchange*, *Knowledge Transfer*. For a visual representation of this division, see 2.1.

Scientometrics. To develop a framework capturing biases of rankings (**RQ1**) and develop new indicators (**RQ2** and **RQ3**), we use two different dumps of the *Microsoft Academic Graph*(MAG) [213]. For answering **RQ1** and **RQ2**, we use the first version of this database, that we call `MAG16`. This version was released for the KDD–cup of 2016, a computer science competition linked to a prestigious conference on knowledge discovery and data mining[2]. We use the second version of the MAG (`MAG17`) for answering **RQ3**. This second version was released by the Open Academic Society[3] in 2017. Both versions of the MAG contain more than 100 million publications, and for each publication, we have a unique publication identifier matched to its title, publication date, journal (or venue), and many other pieces of information. The main difference between these two versions of the MAG is that name disambiguation was provided in the first version, but not in the second. We provide a complete description of `MAG16` and `MAG17` respectively in Sect. 2.1.1A and in Sect. 2.1.1B.

[2] `www.kdd.org`
[3] `https://www.openacademic.ai/`

Collaboration and knowledge exchange. In **RQ4** and **RQ5**, we address the problem of quantifying knowledge exchange occurring during collaborations. In particular, we focus on collaborations observed in two distinct domains: economic and academic. In the *economic domain*, we look at alliances established for Research & Development between firms. These alliances are listed in the SDC Platinum dataset that we describe in Sect. 2.1.1C. To capture the knowledge of firms, we reconstruct their patent portfolios using the patent database of the National Bureau of Economic Research (NBER) (see Sect. 2.1.1D). In the *academic domain*, we look at scientists co-authoring papers in physics journals published by the American Physical Society (APS). The APS provides a dump of its data that contains more than 450 000 papers and dates back to 1893. For these papers, we have many important metadata including title, authors and PACS[4] codes (see Sect. 2.1.1E). Note that PACS codes are extremely relevant as they allow us to assign papers to different research fields. Assuming that papers in different research fields contain different types of knowledge, PACS codes proxy the different knowledge of the scientists writing these papers.

Knowledge Transfer. Knowledge is not only contained in knowledge artifacts, such as patents and papers, but also in human beings. Hence, to understand where and why knowledge moves, we analyze and model the geographical mobility patterns of scientists (**RQ6** and **RQ7**). To do this, we reconstruct the career trajectories of scientists using two datasets extracted from MEDLINE, the largest bibliographic database in the life sciences. The first dataset is `Author-ity` that provides disambiguated author names, linking them to their respective papers [229]. The second dataset is `MapAffil` that provides for each paper and authors disambiguated city names of the listed affiliations [228]. We present these datasets in Sect. 2.1.1F. Additionally, we analyze the mobility patterns of scientists at the affiliation level using the `MAG16` data. We used this data set at the affiliation level as it covers more disciplines compared to MEDLINE.

[4]Physics and Astronomy Classification Scheme (PACS).

2.1.1 Full list of dataset and description

A – *Microsoft Academic Graph* (`MAG16`): KDD–cup version

The KDD Cup is a yearly competition linked to the most prestigious computer science conference about Knowledge Discovery and Data Mining (KDD). For this competition in 2016, a dump of the *Microsoft Academic Graph* (MAG) was released [213]. It contains more than 126 million of publications and more than 467 million citations. Each publication is also endowed with various properties such as unique ID, publication date, title, and journal ID.

Among the primary interests of the community of computer scientists organizing the KDD Cup, there are the technical challenges related to web-scale data collection and aggregation. For this reason, the released data for the KDD Cup 2016 went through only basic processing[5]. Hence, we further pre-process the data to remove from the analysis papers with incomplete information (more details are in Appendix A). With our pre-processing, we obtain $N = 18\,193\,082$ unique publications and $E = 109\,719\,182$ citations.

Additionally, to verify the data quality of this dump, we address the following question: to which extent is the KDD dump of the MAG in accordance with the most updated online version of the MAG[6]? For this comparison, we have randomly sample 50 000 papers that are approximately 0.1% of the total number of papers. First, we have matched the hexadecimal publication ID present in the data released for the KDD Cup to the `int64`publication ID present in the on-line version of the MAG. For 50 papers, we have manually verified that the paper IDs were the same in the two datasets, albeit represented in different formats. Then, using the Academic Knowledge Api[7], we have downloaded the number of citations for each sampled paper.

For the 2.3% of the sampled papers, we were not able to find their corresponding paper in the online MAG. For 50 unmatched papers, we found out that these were papers with duplicates in the KDD Cup data, i.e., these papers are present in the KDD data with two distinct IDs, and only one of the two IDs is

[5] `https://kddcup2016.azurewebsites.net/Data`
[6] We have performed this analysis during February 2017.
[7] `https://www.microsoft.com/cognitive-services/en-us/academic-knowledge-api`

present in the MAG. For the matched papers (97.7% of the sample), we compared the number of citations reported in the KDD dump of the MAG with the number of citations reported on the online version of the MAG. Since the online MAG covers more years than the KDD data, in the absence of noise, we expect the number of citations in the KDD data to be smaller than or equal to the number of citations reported in the online MAG. We find that the two citation counts are highly correlated (Fig. 2.2), and only the 2.2% of the sampled papers have more citations in the KDD data compared to the online version of the MAG. This sets a lower boundary for the error: the percentage of papers with the wrong number of citations is 2.2%.

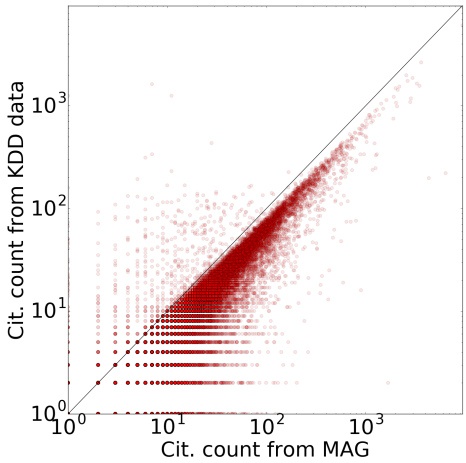


Figure 2.2: Scatter plot of the citation counts reported in the data released for the KDD Cup 2016 and from the online version of the MAG (02/2017).

To summarize, after our filtering procedure, we find that the data released for KDD Cup 2016 has about 2.3% of papers with duplicates. In addition, at least 2.2% of the matched papers have errors in their citation count. This means that we have correct citation information for about 95.6% of the analyzed papers.

B – *Microsoft Academic Graph* (MAG17): Open Academic Society version

A vast group of public and private institutions compose the Open Academic Society. This society aims to create a shared, open, and expanding knowledge graph of research, containing education-focused entities and relationships. Indeed, they provide a free download of two large bibliographic databases: MAG (166 192 182 papers) and AMiner (154 771 162) both updated to 2017. Additionally, they have generated 64, 639, 608 matching relations between the two databases.

We have downloaded the 166 192 182 papers from the MAG. The downloaded data was provided in 9 zip files each of them containing papers information as JSON object. The main disadvantage of this version of the MAG compared to the previous one is that the authors' names are not disambiguated, i.e., each paper has an authors' list containing authors' names and not unique authors' IDs. The main advantages are two: MAG17 is more updated and its information can be cross-checked using the AMiner dataset.

For addressing **RQ2**, we do not need disambiguated authors' names. Hence we use this second version of MAG as it is the more updated compared to the first version.

C – SDC Platinum (SDC)

SDC Platinum database[8] contains data about approximately 672 000 announced alliances from all countries between 1984 and 2009 with daily resolution. The economic actors participating in these alliances are of several types, e.g., manufacturing firms and universities, but for simplicity, we address all of them as *firms*. Each firm listed in the data set is associated with a SIC (Standard Industrial Classification) code that allows us to assign its corresponding industrial sector unambiguously. Further, the purpose of each alliance is characterized by various flags, e.g., manufacturing, licensing, research and development (R&D). We restrict ourselves to all alliances with the flag "R&D", which gives

[8] http://thomsonreuters.com/sdc-platinum/

us 14 829 alliances connecting 14 561 firms. The number of partners involved in each alliance can vary. In most cases, it is two but can also be three or higher. An interesting comparison between the SDC data and other alliance databases is provided by [198]. The author finds that most alliances listed in a database are not present in the others, and hence, she shows that each database should be considered a sample of the collaborations established in the real-world. Also, she finds that the SDC had the broader coverage of alliances across industrial sectors and included many non-OECD alliances. Hence, by using SDC, we are using (to our knowledge) the more complete databases listing alliances spanning from different countries and sectors. At the same time, we know that this database only represents a sample of all the established alliances.

D – Patents from the U.S.A. National Bureau of Economic Research (NBER)

The Patent Citations Data by the NBER contains about three million unique patents granted in the U.S.A. between 1976 to 2006.[9] For each patent, we have various types of information such as the assignees, application year, granted year, and IPC codes. We are particularly interested in the assignee information and the Industrial Patent Classification (IPC) codes. Using the assignee information, we can match the firms listed in the SDC Platinum dataset, and hence, we reconstruct for these firms their patent portfolios. Using the IPC codes of patents, we proxy the knowledge of the firms depending on their patents. We provide more details on how we do this in Sect. 5.2.2.

E – American Physical Society (APS)

This dataset contains over 450 000 papers published in any APS journal, namely Physical Review Letters, Reviews of Modern Physics, and all Physical Review journals, between 1893 and 2009 (116 years)[10]. For each publication listed in this dataset, we have various metadata information, such as the DOI, authors

[9] https://sites.google.com/site/patentdataproject/Home
[10] http://www.aps.org/

list, PACS codes of the papers, and publication date. One limitation of the reported metadata is that authors are identified by their names. Thus, in order to really make use of the APS data set to study the co-authorship activities, we need to disambiguate authors' names. Sinatra *et al.* [212] provides a list of author names disambiguated together with the DOIs of their co-authored paper. This data set contains more than 230 000 disambiguated authors names and the total number of distinct papers in this second data set is 425 118. By matching paper DOIs from the [212] and the original data, we assign to each author a set of papers that he/she has co-authored together with the PACS codes of these papers. Note that PACS codes were introduced only for papers published after 1975, and hence, we retain information about authors authoring papers after 1975.

F – MEDLINE: `Author-ity` and `MapAffil`

MEDLINE is the largest bibliographic database in the life sciences made available by the U.S. National Library of Medicine (NLM). It covers papers published from 1966 to present from more than 5 200 journals in about 40 languages. The first subject scope of this database is biomedicine and health. At the same time, these are broadly defined, and hence, MEDLINE includes papers belonging to the areas of behavioral sciences, chemical sciences, bioengineering, environmental science, marine biology, biophysics, plant, and animal science.

From this database, we use two datasets extracted by *Torvik Research Group*[11]. In particular, we use `Author-ity` [229] and `MapAffil` [228]. `Author-ity` contains *disambiguated author names* linked to their papers from 1966 to 2009. This dataset allows reconstructing for each author his/her list of publications. `MapAffil` lists for each MEDLINE paper and each author the *disambiguated city names* of the listed affiliation (37 396 671 city-name instances). It further gives a unique identifier as well as the geo-coordinates of each city. This second dataset covers publications from 1966 to 2015. With these the two datasets, we can extract for each given author all the cities of her affiliation and the dates of the associated publications. Combining these two sources of information

[11] http://abel.lis.illinois.edu/

about geo-coordinates and time allows us to construct the "career trajectory" of those scientists that have published in that time. These trajectories are the sequences of cities where scientists have worked during their scientific careers (as witnessed by their publications). An example of such a career trajectory is given in Table 2.1. The merged dataset contains in total the career trajectories of $N = 3\,740\,187$ scientists, which were active in the period between 1966 and 2009, traversing $M = 5\,485$ unique cities.

	Year	Affiliation City	PubMed ID
1	2003	Stony Brook, NY, USA	12703729
2	2003	Stony Brook, NY, USA	12595470
3	2005	Kansas City, KS, USA	15936007
4	2005	Stony Brook, NY, USA	15791955
5	2005	Stony Brook, NY, USA	15944300
6	2005	Milwaukee, WI, USA	16299285
7	2007	Milwaukee, WI, USA	17311921
8	2007	Milwaukee, WI, USA	17490406
9	2008	Boston, MA, USA	18566416
10	2008	Stony Brook, NY, USA	18591234

Table 2.1: Example of career trajectory of a specific author (Zhang Y.). For each record we have the year of publication, the city of the affiliation and the PubMed ID identifying the paper. (The PubMed ID is the unique identifier of the paper within the MEDLINE corpus)

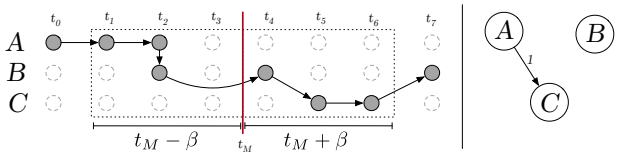


Figure 2.3: Illustration of the procedure used to extract scientists' movements.

Formally we denote a career trajectory of an author $i \in N$ as a sequence p_i, for example $p_i = \{A_{t_0}, A_{t_1}, A_{t_2}, B_{t_2}, B_{t_4}, C_{t_5}, C_{t_6}, B_{t_7}\}$. A denotes the city as defined by its geo-location $R_A = (X, Y)$ where X gives the latitude and Y the

longitude according to the data from `MapAffil`. The subscript t_0 refers to the time measured in years that an author i was based in the respective city. An illustration is shown in Figure 2.3. Note that due to the time resolution of one year, an author may have multiple publications as well as multiple locations in the same year. This can be seen in Figure 2.3 at t_2 where an author is observed in city A and in city B.

G – SCImago

The SCImago Journal & Country Rank[12] is an open access portal that provides indicators for journals and country using data contained in Scopus[13]. From this portal, we have crawled the two years journal impact factors. The two years journal impact factor is the average number of citations received by documents belonging to a specific journal published in a two years time window. Precisely it is computed by counting the "citations received in year X to documents published in years X-1 and X-2"[14]. We have aggregated this score at scientists level and then, at the city level. By this, we have obtained a fitness indicator for scientists and cities. We will use this indicator when modeling scientists' career trajectories in Chap. 9.

2.2 Network theory

When modeling and analyzing a system, we will often adopt a network perspective. With this perspective, we abstract from the details of the analyzed system, and we retain only information about its elements represented as nodes and their interactions represented as links. This allows us to have a simplified but yet interesting representation of the analyzed system. Indeed, a network perspective was successfully applied to different problems ranging from information retrieval, statistical physics, ecology, and sociology. Models and analysis based on a network perspective go under the name of network theory.

[12] `https://www.scimagojr.com/index.php`
[13] `https://www.scopus.com/`
[14] `https://www.scimagojr.com/help.php`

2.2.1 Networks and their basic properties

In network theory, a network is an ordered pair $G = (V, E)$ where $V = \{v_1, ..., v_N\}$ is a set of nodes and $E = \{(v_i, v_j), ..., (v_k, v_l) : v_i, v_j, v_k, v_l \in V\}$ is a set of links connecting the nodes in V. A critical property of a network is that it can be either directed or undirected. A network is directed when a link $(v_i, v_j) \neq (v_j, v_i)$. This type of network is used to represent systems where elements have directed interactions. For example, citation data are often represented as directed citation networks: a document-i citing a document-j is represented as a node-i with a direct link towards node-j. An undirected network is a network where a link $(v_i, v_j) = (v_j, v_i)$. This type of network is used to study systems where interactions among the systems' elements do not have a direction. An example is a social network, like Facebook: when two users declare to be friends, we do not have a direction in their friendship link.

Both directed and undirected networks can be represented by using an adjacency matrix, $\mathbf{A}$. The adjacency matrix of a network $G = (V, E)$ is a matrix where its element A_{ij} represents the number of links $(v_i, v_j) \in E$. Note that if a network is undirected, its adjacency matrix is symmetric, i.e., $\mathbf{A} = \mathbf{A}^T$. For a directed network, this is often not true.

When the possible values of $\mathbf{A}$ are either one or zero, then this adjacency matrix is representing an unweighted network. An example of an unweighted network is a social network composed of friendship links: two users i, j are either friends ($(v_i, v_j) \in E \Rightarrow A_{ij} = 1$) or not ($(v_i, v_j) \notin E \Rightarrow A_{ij} = 0$). When the possible values of the elements of the adjacency matrix are real numbers, the network is usually called a weighted network. An example of a weighted network is the interbank network where nodes are bank and links represent the credits or debits between banks. Note that to include the weights of each link, the edge set E changes from being a set of tuple (v_i, v_j) to a set of triplets (v_i, v_j, A_{ij}).

Path, distance and diameter. Given a network $G(V, E)$, two nodes $v_i, v_j \in V$ can be directly connected by a link $(v_i, v_j) \in E$ or by a sequence of links, e.g. $((v_i, v_k), (v_k, v_j))$ both belonging to E. A sequence of links on a network is called *path* and is usually represented by the traversed nodes. For example,

the sequence of links $((v_i, v_k), (v_k, v_j))$ is written as the path $(v_i \to v_k \to v_j)$. A path made of n links has length n. The shortest path length between two nodes v_i, v_j is the (network or topological) *distance* between these two nodes, $d(v_i, v_j)$. If two nodes do not have a path connecting them, they are considered disconnected, and they have infinite distance. The maximum distance between nodes is the *diameter* of the network, i.e., it is the length of the longest among the shortest paths. Note that if a network is undirected network distances are symmetric, meaning that $d(v_i, v_j) = d(v_j, v_i) \, \forall i, j$. While on a directed network $d(v_i, v_j)$ can be different from $d(v_j, v_i)$.

In the next section, we present two types of measures commonly used in network theory. The first type of measures provide scores to individual nodes and allow us to identify the critical nodes in the network. These measures are often named *centrality measures*. The second type of measures are called *clustering algorithms*. These are algorithms that group nodes depending on the network structure.

2.2.2 Centrality measures

Degree. In a un-directed network G with adjacency matrix $\mathbf{A}$, the degree of a node i is defined as $k_i = \sum_l A_{il}$. In a directed network G with adjacency matrix $\mathbf{A}$, we define in-degree of node i as $k_i^{in} = \sum_l A_{li}$ and its out-degree as $k_i^{out} = \sum_l A_{il}$. In other words, the in-degree of a node i is the sum of the incoming links to i, while the out-degree is the sum of the outgoing links.

Betweenness. Nodes are directly connected by links and indirectly connected by paths. Hence, nodes lying on paths are critical as they connect nodes that do not have direct links. The betweenness centrality measures the importance of a node by counting the shortest paths to which it belongs. In formula, given a network $G(V, E)$, the betweenness centrality of a node $v \in V$ is

$$b(v) = \sum_{s \neq v \neq q} \frac{\sigma_{sq}(v)}{\sigma_{sq}} \tag{2.1}$$

where σ_{sq} is the number of shortest paths between nodes s and q both belonging to V and $\sigma_{sq}(v)$ is the number of shortest paths between nodes s and q traversing v. Note that in weighted networks, the definition of betweenness centrality has to be changed to account for the weights of the links.

Local clustering coefficient. Given a network $G(V, E)$ and a focal node $v \in V$, the local clustering coefficient measures how connected are the neighbors of v. Precisely, it is the number of links between the neighbors of v, divided by the number of possible links between the neighbors. In formula, the local clustering coefficient of v is

$$c(v) = \frac{2e}{k(k-1)} \tag{2.2}$$

where k is the degree of v and $e = |(w, y) \in E : (w, v) \in E \wedge (y, v) \in E|$. By taking the average on all the nodes of their local clustering coefficient, we obtain the average clustering coefficient for the network.

Note that here we have covered only three fundamental nodes centrality measures. There exist a large number of centrality measures that capture very different properties of the nodes. For an almost complete list of these measures and their interpretation, see [144].

2.2.3 Clustering algorithms

Many real-world systems are organized in groups. For example, inside ecosystems, we often find groups of pollinator and plant species strongly interacting with each other [156]. Similarly, in social organizations like firms, we find individuals divided into teams: members of the same team frequently interact among each other, while interactions across teams are rarer. In network theory, these groups or teams are named *communities*, and their identification is performed using clustering algorithms. Precisely, a clustering algorithm identifies groups of nodes that are called *modules*, and the set of identified modules is called *clustering*. A good clustering algorithm should identify modules that match the communities present in a system.

Note that there exists a wide variety of clustering algorithms. Probably the number of clustering algorithms is even larger than the number of network measures. This is because the clustering problem is not well defined. A clustering algorithm should identify modules that contain a set of nodes "more connected" inside the module than outside. Depending on the system under analysis, the "more connected" can have different definitions. In other words, depending on the system analyzed, the meaning and definition of a community is different.

We present only two algorithms that are often used: the `Louvain algorithm` and `Infomap` [24, 188]. We concentrate on these two as both have been demonstrated to work very well on both synthetic [112] and real-world data [60]. Additionally, the `Louvain algorithm` is based on *modularity* [150], which is one of the most used measures to define "more connected" nodes. `Infomap`, instead, defines "more connected" nodes depending on a random walk process on the network. Such a process is constructed by normalizing the network adjacency matrix to create a *transitions matrix*. With this matrix, we can define a Markov chain where nodes are possible states, and the elements of the transition matrix are transition probabilities between the states. Using such a process, `Infomap` identifies clusters on networks. In the next two paragraphs, we describe with more details both clustering algorithms.

`Luoivan algorithm.` In order to capture sets of more connected nodes, many clustering algorithms rely on the modularity score [150]. This score is computed by taking the difference between the fraction of *observed* links that fall within the same modules and the fraction of links that are *expected* to fall within the same modules. Given a network $G(V, E)$ and a clustering of this network $\mathcal{C}$, the fraction of *observed* links falling within the same modules can be computed by summing over the elements of the adjacency matrix:

$$\frac{1}{2m} \sum_{v,w} A_{vw} \delta_{c_v, c_w} \tag{2.3}$$

where $m = |E|$, δ_{c_v, c_w} is the Kronecker delta (if $i = j$, $\delta_{i,j}$ is equal to 1; else it is equal to 0) and c_v is the cluster of node v. The fraction of *expected* links to fall

within the same modules depends on the null model used. Usually, researchers use the configuration model, that is a random network model that preserves the degree of the nodes. Hence, the number of links expected between two nodes v and w is $\frac{k_v k_w}{2m}$ and the modularity score is

$$Q = \frac{1}{2m} \sum_{v,w} \left(A_{v,w} - \frac{k_v k_w}{2m} \right) \delta_{c_v, c_w} \tag{2.4}$$

The `Louvain algorithm` maximizes the modularity score Q following three steps. *First*, it assigns every node to a separate module, i.e., each module contains a single node. *Second*, it computes the changes in modularity when moving a node v to the modules of its neighbors. Then, the algorithm assigns v to the module, causing the maximum positive change. This second step is iterated over all the nodes more than once until no movements can increase the modularity score. *Third*, the algorithm makes a size reduction [9]: each module becomes a (macro-)node, and the links across modules become links among the (macro-)nodes, while the links within the modules become self-loops. On this reduced network, it performs again the above steps until the modularity cannot be increased anymore. For a more detailed description of this algorithm, see the original paper [24].

The advantage of the `Louvain algorithm` is that it is based on the modularity score. This is a well-accepted measure of connectedness and is mathematically well-defined. However, this measure has also some limitations that affect the `Louvain algorithm`. For example, in a big network with many small communities, the modularity score increases when merging these smaller communities in one module [58]. This limitation is called *resolution limit* and defines the minimum size of the communities detect by `Louvain algorithm`. Additionally, the `Louvain algorithm` is dependent on the null model used to compute the modularity. In other words, it is dependent on the assumption that the expected number of links between any nodes v and w is $k_v k_w / 2m$. Hence, the `Louvain algorithm` is dependent on the null model assumed to form the network.

`Infomap`. This algorithm is based on two concepts: random walk and compression. The random walk is a stochastic process where a walker is assumed to move randomly in a continuous or discrete space. In network theory, the random walker moves on a network from one node to another by following links. `Infomap` assumes that "more connected" nodes are those nodes that are "more frequently visited" one after the other by a random walker. In other words, two nodes v and w belong to the same module when it is easy to reach the node w by randomly following links starting from v. To quantify which nodes are "more frequently visited", `Infomap` compresses the sequences of the visited nodes. This compression is achieved by first encoding these sequences using the *Huffman coding*. Then, the description length of the encoded sequences is minimized using the *map equation*. The details about the *Huffman coding* and the *map equation* can be found in [188].

`Infomap` has a totally different approach to create modules compared to algorithms based on modularity maximization, like the `Louvain algorithm`. From a technical point of view, `Infomap` never computes the modularity score. It only simulates a random walk process on the network and then, clusters nodes depending on how frequently these nodes are visited one after the other. From a conceptual point of view, `Infomap` creates modules by analyzing how the network structure influences its functionality. Indeed, by assuming that a network carries a flow (represented by the random walk process), it clusters nodes where this flow is more stagnant. This allows us to detect communities depending on the functioning, rather than the formation of the network.

In the next chapters, we will use `Infomap` to detect communities in the network for two main reasons. First, we will be dealing mostly with large networks with lots of links, but we will not have information about the size of the communities on these networks. Hence, the resolution limit of the modularity score might be an issue that we would need to correct. Second, in Chap. 4, we will study journals depending on the knowledge traversing them. We will do this by assuming that knowledge flows from cited paper to citing papers, i.e., that knowledge flows in the opposite direction of citations. Then, `Infomap` will perfectly fit this study as it detects communities by identifying where the flow on a network is stagnant. In our case, this will mean that journals will be assigned to groups where the knowledge is more similar.

As a final remark, let us remind that the correct clustering algorithm to be used depends on the system and on the analysis that a researcher wishes to perform. Various review works can help to decide which algorithms should be used. One of the most comprehensive and educative review work is [57], while a shorter review with an engaging introduction for physicists is [167]. For a comparison of clustering algorithms based on performance (accuracy and speed), see [45], while for a comparison based on the definition of communities, see [41].

2.3 Scientometric indicators

To quantitatively study the relation between knowledge artifacts, like papers and patents, we rely on scientometric indicators. In this thesis, we will focus on scientometric indicators based on citation analysis and a network perspective. This type of indicators has an established role in determining the scientific impact of scientists and organizations [94]. At the same time, their use is also highly disputed [1]. We dedicate the following section to introduce citation-based indicators frequently used to detect papers' impact and journals' similarity. Here, we only introduce their definition and usage, but we will not analyze them. Their analysis will be performed in Chap. 3 and Chap. 4. In these chapters, we will discuss their shortcomings and propose new indicators to overcome these shortcomings.

2.3.1 Indicators of paper impact

We now define four indicators used to detect paper impact: citation count c, relative citation count c^f, PageRank score p, age-rescaled PageRank score $R^A(p)$.

Citation count, c. The citation count c_i of node i is simply the number of citations received by paper i. In terms of the citation network's adjacency matrix $\mathbf{A}$ (in a directed network, $A_{ij} = 1$ if node i points to node j, otherwise $A_{ij} = 0$), we can express the citation count as $c_i = \sum_j A_{ji}$.

Relative citation count, c^f. To overcome citation count's bias by paper age and academic field, [179] defined the *relative citation count* c_i^f of paper i as $c_i^f := c_i/\mu_i^Y(c)$, where $\mu_i^Y(c)$ denotes the mean citation count for papers published in the same field and year as paper i.

PageRank, p. Citation count and indicators built on it share a notable limitation: the citations a paper receives are all counted the same, regardless of the importance of the citing paper. A possible way to overcome this limitation – recognized already in the 70s in the scientometric community [164] – is to take into account the whole structure of the paper-paper citation network. In this spirit, eigenvector-based indicators take as input the citation network's adjacency matrix $\mathbf{A}$. This class of indicators has been applied in various research domains, including scientometrics [23, 164], Web information retrieval [31, 106], social science [25, 104] – see [53, 61, 76] for a review. Among these indicators, we focus on Google's PageRank score [31]. This score was initially devised to rank webpages in the World Wide Web and has attracted lots of interest in the scientometric community. The rationale behind its application to citation networks is that citations coming from influential papers should count more than citations from unknown papers.

The PageRank scores of papers are usually written in a vector $\mathbf{p}$ defined by the following equation

$$\mathbf{p} = \alpha\,\mathbf{P}\,\mathbf{p} + (1-\alpha)\mathbf{v}, \tag{2.5}$$

where α is a parameter of the algorithm (called damping factor), $\mathbf{P}$ is the random-walk transition matrix with elements $P_{ij} = A_{ij}/k_j^{out}$, $k_j^{out} = \sum_l A_{lj}$ is the number of references in paper j, and $\mathbf{v}$ is a uniform teleportation vector with elements $v_i = 1/N$ for all papers i. Eq. (2.5) can be interpreted as the stationary equation of a stochastic process on the citation network. In this process, a random walker is placed on each paper and the walker either follows a citation edge with probability α or jumps to a randomly chosen paper with probability $1-\alpha$. When the number of walkers on each paper reaches a stationary value, the PageRank score of a paper i is the fraction of walkers on this paper. There is no universal criterion to choose the value of the damping factor α. In agreement with [36], we set $\alpha = 0.5$ which corresponds to a

random walker covering paths of length two before teleporting to a random node. [36] argue that the choice $\alpha = 0.5$ better reflects the actual surfing behavior of researchers than the commonly used value $\alpha = 0.85$.

PageRank is based on a static, time-aggregated perspective of the considered network. Such a perspective is limiting for the analysis of evolving networks [129, 204] and indeed, PageRank has been found to be biased in favor of old papers [36, 129, 130, 133].

Age-rescaled PageRank, $R^A(p)$. To suppress the age bias of PageRank, Mariani *et al.* [130] proposed to rescale the PageRank score. Assuming that the papers are ordered by older to younger, one computes the mean value $\mu_i^A(p)$ and the standard deviation $\sigma_i^A(p)$ of PageRank scores over Δ_p papers around paper i, i.e. $j \in [i - \Delta_p/2, i + \Delta_p/2]$. Consequently, the rescaled PageRank score $R_i^A(p)$ of paper i is defined as

$$R_i^A(p) = \frac{p_i - \mu_i^A(p)}{\sigma_i^A(p)}. \tag{2.6}$$

The authors of [130] applied rescaled PageRank to the network of physics papers to show that the resulting ranking is not biased by paper age. This allows us to identify seminal publication much earlier than rankings by indicators that are biased against recent papers. In the following, we set $\Delta_p = 1000$ as in [130].

2.3.2 Indicators of journal similarity

We now define two citation-based indicators: bibliographic coupling and cosine similarity based on co-citations. These are commonly used to compute the similarity scores between papers, authors, or journals. Since in this thesis we are particularly interested to compare journals (see Chap. 4), we provide *first* their initial definition (used to compare papers) and *then*, their extension to the journal case.

Bibliographic coupling. This measure was introduced in [105] to group technical and scientific papers automatically. It is based on the concept of *bibliographic coupling unit*, i.e., a paper cited by two different papers. Then, the bibliographic coupling (strength) between two papers is defined as the number of coupling units that these papers share. Given an adjacency matrix **A** of the citation network at the paper level, the bibliographic coupling between two papers p_i and p_j is

$$B_c(p_i, p_j) = \sum_k A_{ik} A_{jk} \tag{2.7}$$

where the element A_{ik} is 1 if the paper i cites paper k, and zero otherwise. This measure has a limitation: it is static at the paper level as it is dependent on the older papers co-cited by i and j. In other words, it does account for the evolution of knowledge as the similarity between papers does not change when new papers are produced. However, this limitation is overcome when we aggregate citation at the journal level. Journals continuously publish new papers, and hence, also their out-going citation change together with their bibliographic coupling. Note also that bibliographic coupling captures the "outward" similarity between two papers or journals as it is dependent on their out-going citations. One of the first applications of bibliographic coupling to journals was done by Small and Koenig [216]. Here, the authors also normalize this measure to account for journal-size effects. The bibliographic coupling between two journals, C and D, is:

$$B_c(C, D) = \frac{\sum_{p_k \notin C,D} \left(\sum_{p_i \in C} \sum_{p_j \in D} A_{ik} A_{jk} \right)}{\left(\sum_{p_i \in C} \sum_{p_k \notin C} A_{ik} \right) \left(\sum_{p_j \in D} \sum_{p_k \notin D} A_{jk} \right)} \ . \tag{2.8}$$

Cosine similarity based on co-citation. The co-citation similarity measure was introduced independently by Small [215] and Marshakova [132]. This measure captures the similarity between papers depending on their "in-coming" citations. Two papers p_i and p_j are similar if other papers cite both p_i and p_j. When two papers are just published, their similarity is 0; as time goes by, their similarity can increase. For this reason, co-citation similarity is by

construction evolving in time. Given the adjacency matrix $\mathbf{A}$ of a citation network, the co-citation similarity between two papers p_i and p_j is

$$C(p_i, p_j) = \sum_k A_{ki} A_{kj} \tag{2.9}$$

where the element A_{ki} is 1 if the paper k cites paper i, zero otherwise. White and Griffith [254] extended this measure to capture author similarity, while McCain [135] used it to group journals. Their idea was to aggregate the co-citation scores of papers belonging to the same authors or journals.

To account for the different sizes of journals, we can normalize the co-citation similarity in different ways. Here, we consider its "cosine" normalization [235] that is (probably) the most popular one [52]. Then, based on co-citation, the cosine similarity of two journals, B and D, is:

$$C_c(B, D) = \frac{\sum_{L \in J \setminus \{B,D\}} \sum_{p_k \in L} \left(\sum_{p_i \in B} \sum_{p_j \in D} A_{ki} A_{kj} \right)}{\sqrt{\sum_{L \in J \setminus B} \left(\sum_{p_i \in B} \sum_{p_k \in L} A_{ki} \right)^2 \cdot \sum_{L \in J \setminus D} \left(\sum_{p_j \in D} \sum_{p_k \in L} A_{kj} \right)^2}} \tag{2.10}$$

2.4 Agent-Based and Network models

In this thesis, we take a network perspective to analyze socio-technical systems. Also, we reproduce their properties using agent-based models. We have chosen this modeling approach out of many possible candidates. We now motivate such a decision by discussing four popular models for longitudinal relational data: Temporal Exponential Random Graph Models (TERGM) [89], Stochastic Actor-Oriented Based Model (SAOM) [217], Relational Event Models (REM) [33], and Data Driven Agent-Based Models. The following discussion wil not be sufficient to understand all the details of these models. At the same time, it should be sufficient to understand why we have chosen data driven agent-based models instead of the other models.

We start our discussion by describing two of the most popular models for social scientists to analyze longitudinal relational data, i.e., an ordered se-

quence of networks $(G^t)_{t=1}^T$. The first model is the Temporal Exponential Random Graph Models (TERGM) and the second model is the Stochastic Actor-Oriented Based Model (SAOM). Both models are defined using the famous Exponential Random Graph Model (ERGM) [252]. An ERGM is fully described by the probability function:

$$P(G|\vec{\theta}) = \frac{e^{\vec{\theta}\cdot\vec{h}(G)}}{Z(\vec{\theta},\vec{h})} \tag{2.11}$$

where G is the observed network, $\vec{\theta}$ are the parameters weighting the importance of the effects/statistics $\vec{h}(G)$ calculated on the network G, and Z is the partition function. This function is defined as $Z(\vec{\theta},\vec{h}) = \sum_{W\in\mathcal{G}} e^{\vec{\theta}\cdot\vec{h}(W)}$ with $\mathcal{G}$ representing the set of all possible network with the same number of nodes as G. Note that each component of $\vec{h}(G)$ represents one of the dependent variables of the ERGM and these can be both endogenous and exogenous. The former variables can be captured by the network representation of the system, while the latter cannot.

2.4.1 Temporal ERGM (TERGM)

Description. TERGM extends ERGM by defining the probability to observe a network to be conditional to previous observations of the network. In other words, given a network G^t at time t, the probability to observe G^t is

$$P(G^t|G^{t-\tau},...,G^{t-1},\vec{\theta}) = \frac{e^{\vec{\theta}\cdot\vec{h}(G^{t-\tau},...,G^{t-1},G^t,\vec{\theta})}}{Z(G^{t-\tau},...,G^{t-1},G^t,\vec{\theta},\vec{h})} \tag{2.12}$$

where τ is the memory of the system, i.e., how far back in the past the sources of influence on the observed network are. Indeed, the statistics $\vec{h}$ is not only dependent on the observed network, but also on its previous observations $G^{t-1},...,G^{t-\tau}$ up to time $t-\tau$.

Advantages. We can model sequences of observed networks without the need to assume temporal independence. Given a sequences of network $(G^t)_{i=1}^T$, the probability of observing the last network G^T is

$$P(G^T|(G^t)_{t=1}^{T-1}, \vec{\theta}) = \prod_{t=1}^{T-\tau} P(G^{t+\tau}|G^t, ..., G^{t+\tau-1}, \vec{\theta}) \tag{2.13}$$

The temporal dependence is introduced by the effects in $\vec{h}$ that capture network properties linked to time. For example, given a sequence of unweighted networks $(G^i)_{i=1}^t$, then the stability of links between nodes is captured by the memory effect $\sum_v \sum_w A_{vw}^t A_{vw}^{t-1} + (1 - A_{vw}^t)(1 - A_{vw}^{t-1})$ where $\mathbf{A}^i$ is the adjacency matrix of the network G^i. In this discussion, we restrict our attention to data represented using unweighted and undirected networks.

Disadvantages. The disadvantages of the TERGM are mainly of two kinds. The first kind comes from the ERGM that we need to define a TERGM. To use a TERGM, we have to fit an ERGM at every time step correctly. ERGMs are known to suffer from many problems such as their inability to deal with co-linearity among the effects. Additionally, they suffer from computational issues when dealing with large networks. Indeed, even after a long computation time, the networks resulting from fitting of ERGMs can be far away from the observed networks.

The second kind of disadvantages is that TERGMs do not have a close connection to the data generating process. In the vector $\vec{h}$, we encode different effects that explain how a network at a time t evolves into a new one at a time $t + 1$. Then, we determine the importance of these effects by looking at their corresponding parameters in $\vec{\theta}$. These are obtained by a fitting procedure consisting of Monte Carlo simulations where edges are added and removed at random using Eq. (2.12). In other words, the TERGM does not reproduce realistic sequences of choices made by the agents(/nodes) present in the system. It only tries to simulate the observed data by adding and removing links.

2.4.2 Stochastic Actor-Oriented Based Models (SAOM)

Description. This is a popular model in social science to study longitudinal relational data and is similar to TERGM as it is also based on Eq. (2.11). However, TERGM and SAOM have many differences, especially in their assumptions and how they reproduce network sequences First of all, SAOM introduces several mini-steps n between two consecutive observations of the network. The number of mini-steps can be large and tend to infinity, and it is usually computed depending on the number of observed changes (i.e., added or removed links) m. During each of this mini-step, nodes – so-called actors – are activated with a certain probability p and they add or remove links depending on an *objective function* f. Usually the activation probability p is computed such that the number of mini-steps n matches the number of changes observed. As a simple example, consider to observe N actors and m changes between two consecutive networks, then we can require that $n \cdot (Np) = m$. We refer to the size of a mini-step using ε, such that $t + n\varepsilon = t + 1$.

When an actor i has been activated, she has the opportunity to make a change, i.e., to add or remove a link. The "direction" of the change (adding or removing a link) depends on an *actor-centric* objective function:

$$f_i(\vec{\theta}, G) = \vec{\theta} \cdot \vec{h_i}(G) \tag{2.14}$$

where each element of $\vec{\theta}$ is a parameter corresponding to an effect in $\vec{h_i}$. Note that the vector containing the effects are computed only from the actor perspective. For example, the effect of link stability from the perspective of actor i is $\sum_v A_{iv}^{t+\varepsilon} A_{iv}^{t} + (1 - A_{iv}^{t})(1 - A_{iv}^{t+\varepsilon})$.

The probability to choose and change a link (i, j) is proportional to e^{f_i} with f_i computed on the network with the link (i, j) changed. The formula to calculate the probability to choose and change the link (i, j) during the k^{th} mini-step is

$$Pr(A_{ij}^{t+k\varepsilon} = 1 - A_{ij}^{t+(k-1)\varepsilon}) = \frac{\exp\ (\vec{\theta} \cdot \vec{h_i}(\hat{G}^{t+(k-1)\varepsilon})\ )}{Z_i(G^{t+(k-1)\varepsilon}, \vec{\theta}, \vec{h_i})} \tag{2.15}$$

where $\hat{G}^{t+(k-1)\varepsilon}$ is the network $G^{t+(k-1)\varepsilon}$, but with the link (i, j) changed ($A_{ij}^{t+k\varepsilon} = 1 - A_{ij}^{t+(k-1)\varepsilon}$) and $Z_i(G^{t+(k-1)\varepsilon}, \vec{\theta}, \vec{h_i})$ is the partition function de-

fined by only those network realizations under the direct influence of i. In other words, $Z_i(G, \vec{\theta}, \vec{h_i}) = \sum_{W \in \mathcal{G}_i} e^{\vec{\theta} \cdot \vec{h_i}(W)}$ where $\mathcal{G}_i$ is a set containing G and only those network realizations with an extra or missing link (i, k) with respect to G.

Advantages. By taking an actor (node) perspective, SAOMs allow us to give different influences to nodes during the network evolution. Indeed, the activation probability p can be chosen to be actor-dependent (p_i), and hence, some actors can be chosen with higher or lower probability during a mini-step. By this, we can impose a different (expected) number of activations per agent between time steps. Hence, by defining different activation probabilities p_i, we model actors' heterogeneity in establishing and terminating relations. Such a heterogeneity is a key aspect of real-world actors belonging to social and economic systems. Additionally, SAOMs produce a precise sequence of events to move from one network to another other. This allows us to reproduce the temporal order in social processes.

Disadvantages. Even though SAOMs can reproduce the temporal order in sequences of events, they cannot simulate simultaneous ones. For example, when an individual sends an email to two other individuals, we should not consider this as two dyadic events, but as a multi-party event. This type of event cannot be reproduced using SAOMs. The second disadvantage of SAOMs is that they are based on a rational-choice of the actor to maximize a specific objective function f_i. This choice is performed only considering one single change in the network, but "not long-run change that will result from immediate changes"[117]. This poses significant limitations on the compatibility of rational-choice theory and SAOM. Additionally, SAOM shares the high computational costs of TERGM.

2.4.3 Relational Event Model (REM)

Description. The key element of this model is a *relational event*, also called *action* a. This is defined by a tuple made of a sender i, a receiver (or a set of receivers) j, the action type k and a time t: $a = (i, j, k, t,)$. Given a sequence

of time ordered actions $(a_l)_{l=1}^t$, REM posits that i) actions are conditionally independent[15] and ii) that probability to observe a new action a at time $t+\tau$ is the hazard function of observing a multiplied by the joint probability that no other events occur between t and $t+\tau$. In formula,

$$Pr(a \text{ at time } t+\tau) = h(a_l) \prod_{a' \in \mathcal{A}_t} S(t, t+\tau | a') \tag{2.16}$$

where h is the hazard function, $S(t, t+\tau|a')$ is the survival probability of action a' between t and $t+\tau$, and $\mathcal{A}_t$ is the set of actions that could occur after time t. Moreover, a REM also specifies the subfamilies of h and S. The former is set to be a constant function λ and the latter to be an exponential $e^{-\lambda\tau}$. Note that λ is a constant function for the rate of actions, but it could be dependent on the sender i, receiver(s) j, type of action k, past action history $(a_l)_{l=1}^t$ or other exogenous co-variates X_a:

$$\lambda(i, j, k, (a_l)_{l=1}^t, X_a, \vec{\theta}) = \exp\left(\lambda_0 + \vec{\theta} \cdot \vec{h}(i, j, k, (a_l)_{l=1}^t, X_a)\right) \tag{2.17}$$

where λ_0 is a constant event rate, $\vec{h}$ are relevant statistics for the action rate, and $\vec{\theta}$ are the parameters of the statistics. When fitting a REM, we evaluate those $\vec{\theta}$ that better reproduce the full sequences of observed events $(a_l)_{l=1}^T$, i.e., that provide a λ maximizing $\prod_l^T Pr(a_l \text{ at time } l)$.

Advantages. With REMs, we model events in an "actor-oriented" fashion similar to SAOM. This modeling assumption allows us to keep a more realistic description of the underlying process generating the observed *sequence of events*. Also, a REM takes as input sequences of events and explains them using the statistics contained in $\vec{h}$. In contrast, TERGMs and SAOMs have *sequences of networks* as input, and they only try to model these. Additionally, REMs allow modeling not only dyadic interactions but also multi-agent ones. For example, we can model an individual sending an email to two or more people as a single action.

[15]Given the occurrence of an event C, two events A and B are conditionally independent if, and only if, $P(A \cap B|C) = P(A|C)P(B|C)$.

Disadvantages. The first disadvantage of REM is that they cannot model actors engaging in forward-looking behavior. This limitation depends on the assumption that actions are conditionally independent, i.e., an actor chooses her next action only considering the history of past actions. An actor cannot make strategic decisions depending on actions *not* taken by the other actors. In other words, given a history of past actions, actors "quickly" respond to this and they lack in forward-looking and strategic behavior. From this "quick" response, we also find the second disadvantage of REMs. They need *complete* time-ordered sequences of relational events as input. If we are missing an event in a sequence, we might miss the reason for the next observed action. Data containing complete sequences of events between individual actors are actually rare. Usually, this type of fine-grained data often misses events, so researchers have to understand if the missing data are biasing the results of a REM.

2.4.4 Data Driven Agent-Based Models

Description. An agent-based model (ABM) is a microscopic model of a system. It is typically defined by its agents, their internal dynamics, and their interactions. The agents represent the elements of the system that a researcher wishes to model. Agents are usually endowed with attributes that represent relevant characteristics of the system elements. The agent's internal dynamics and interactions represents how the system elements evolve and interact, and it should match the microscopic interaction laws of the analyzed system. Like any other model, ABMs are dependent on parameters, and these typically specify the intensity of the interactions between the agents.

We define a *data-driven agent-based model* an ABM for which data is used in three different parts of the model: as *input*, for *calibration* and for *validation*. Using data as input means, for example, that both the initialization of the ABM and the distribution of agents' attributes are dependent on data. A model is said to be calibrated against data when data is used to choose an optimal combination of the model parameters. To choose an optimal combination of parameters, typically an ABM first needs to be simulated with different combinations of parameters. Then, the optimal combination of parameters is the combination with which the ABM better reproduces (previously chosen)

observed properties. Finally, the validation of an ABM consists of verifying that the ABM reproduces observed properties that were *not* used as input or during the calibration procedure. These other properties reflect dimensions of the modeled system that were captured by the ABM even though it was not informed about it.

Advantages. ABMs are a general approach to model systems independently of their representation. SAOMs and TERGMs should be used only to model systems under a network perspective, and REMs are designed to model only sequential data. ABMs, instead, are not restricted to model networks or sequential data. Additionally, they allow us to introduce as many agents attributes or refined interaction rules as necessary to capture the analyzed system. This high flexibility places a massive advantage when modeling complex systems.

The focus of ABMs lies in defining agents' attributes and interactions that reproduce macroscopic properties of the analyzed system. For this reason, they are similar to SAOMs or REMs in keeping agents (actors) at the center of the model. At the same time, they also take a statistical approach similar to TERGM to explain the full system. Indeed, the ability of an ABM to reproduce macroscopic properties should not be dependent on the exact order of interactions among the agents. An ABM typically has some randomness inside that allows for a stochastic evolutions of the system. Such an evolution results in different realizations of the analyzed system, but all these evolution should share similar macroscopic properties. By this, it tests for those microscopic details (i.e., agents' attributes and interactions) sufficient to reproduce macroscopic properties.

Data-driven ABMs are typically generative models, i.e., given a set of microscopic interaction rules among the system elements, an ABM can generate the macroscopic properties of the system. This is something that most contemporary data-driven models do not do. For example, machine learning models can discover and capture statistically significant patterns inside some data about a system. However, they typically do not provide a method to re-generate the found patterns, and when they do provide such a method, it is usually based on a transformation of the observed data. Hence, the generation of the found patterns does not come from an understanding of the analyzed system.

Last, but not least, an ABM can be more computationally efficient than SAOM, TERGM, and REM. For every system that we wish to model, we can tailor a new ABM with minimal ingredients. Moreover, the different realizations of modeled system can be easily parallelled as they are independent.

For the present thesis, we are mostly interested in studying how individual actor decisions shape the system in which they are found. For this reason, we decide to use ABMs instead of TERGMs. Also, we do not use SAOMs or REMs as we are not interested in the full details of the interaction sequences. We are only interested in determining those microscopic rules sufficient to reproduce the macroscopic properties of the system. Additionally, we will be dealing with large datasets with often thousands of actors. Simulating systems of this size becomes extremely time-consuming when using SAOM, TERGM, and REM. Hence, we need to develop our programming code tailored to the phenomena that we aim to investigate.

Disadvantages. ABMs are very different from each other even when they are simulating the same system. These differences make it hard or even impossible, a direct comparison between them. These differences often arise from the freedom in mapping the system elements and their dynamics. Indeed, different microscopic rules can be used to explain the same macroscopic property. For example, power-law degree distributions in networks can arise from a preferential attachment (PA) or a coping mechanism. This means that we could develop two different ABMs (one with a PA and one with a coping mechanism) to model the same system property (a power-law degree distribution). However, we have no direct way to decide which model is using the correct microscopic rules. One solution to this problem is to carefully check the theoretical foundations of the proposed rules in each ABM. A second solution is to validate the ABMs against other no-trivial macroscopic features of the system.

The computational efficiency of an ABM comes at the cost of writing its code from scratch. This means that the advantage in the simulation time can be lost as developing and testing the new code is time-consuming. Moreover, usually, a model and its code are prepared to capture one specific system optimally and to be computationally efficient. Hence, the re-usability of a code for modeling a new system is not trivial and often impossible.

2.5 Conclusion

We have introduced the databases and the methods that we will use in the next chapters. Data and methods have been chosen accordingly to the aim of this thesis. Precisely, to study how knowledge is structured, exchanged, and transferred in socio-technical systems, we need diverse databases and a multidisciplinary approach.

To extract relevant information from knowledge repositories, such as patent and bibliographic databases, we need methods coming from *information science*. To interpret and represent the extracted information as one system, we utilize a network perspective, and hence, we need tools from *network theory*. Then, to model and reproduce the reconstructed networks, we will use *agent-based models*.

Information science. We have presented various databases that we will use to answer three different types of questions. The first type of question is about how to develop and test indicators for information science. The second type of question is about how to identify the interplay between collaboration and knowledge exchange. The third type of question is about how knowledge is geographically transferred. For addressing all of these questions, we use bibliographic data: `MAG16`, `MAG17`, `APS`, `NBER`, `MEDLINE`, `SCImago`.[16] To analyze the documents present in these data, we have introduced different scientometric indicators. These indicators are based on citation analysis and are used to measure the impact and similarity of papers and journals.

Network theory. After presenting the data, we have introduced the network perspective that we will use to study knowledge in socio-technical systems. Such a perspective is based on a mathematical object called network, which is an ordered pair of sets. The first set contains nodes that represent the elements of the system; while the second set comprises links that describe the microscopic interactions among the system elements. When tak-

[16] Additionally, for addressing the second type of questions in the R&D domain, we use the `SDC` platinum data listing declared alliances among firms.

ing a network perspective, we combine all the observed interactions into one mathematical object. Then, by studying this object, we can determine the relative importance of the system elements, for example, by using network centrality measures. Additionally, we can identify the mesoscopic structure, such as communities, that are present in the analyzed system.

Data-driven agent-based modeling. We will use ABMs to find the microscopic interaction rules reproducing the systems analyzed in this thesis. We could have used other models, like TERGM, SAOM, and REM. Even though these represent state of the art in network analysis, we have decided not to use them as they do not fit the aim of this thesis. Precisely, we aim at identifying microscopic interaction rules from the actor perspective that explain the macroscopic properties of the systems. To do this, we need ABMs. Note that we will propose *data-driven* ABMs, meaning that data are used as input, and for calibrating and validating the proposed models. Additionally, in our models, we will introduce agents' attributes and their dynamics by considering economic and social science theories. Hence, both attributes and dynamics will be motivated from a theoretical perspective. By all this, our data-driven ABMs will not be simple simulations, but rather a powerful tool to explain systems observed in the real world.

Chapter 3

Quantifying and suppressing ranking bias in citation networks

Summary

In this chapter, we use a large citation data set from Microsoft Academic Graph and a new statistical framework based on the Mahalanobis distance to show that the rankings by well known indicators, including the relative citation count and Google's PageRank score, are significantly biased by paper field and age. Our statistical framework to assess ranking bias allows us to exactly quantify the contributions of each individual field to the overall bias of a given ranking. We propose a general normalization procedure motivated by the z-score which produces much less biased rankings when applied to citation count and PageRank score.[1]

[1]Based on [232]

3.1 Introduction

Paper citation count itself and various quantities derived from it are used as influential indicators of research impact [68, 95]. At the same time, it is well known that the cumulative number of citations received by academic publications strongly depends on paper age and field [205, 240]. Old papers have had more time to acquire citations than recent ones, and their advantage is further enhanced by the preferential attachment mechanism [146, 170]. While heterogeneous paper fitness and paper aging possibly attenuate the advantage of old nodes [136, 251], empirical evidence typically shows that citation count is still biased toward old nodes (see [130, 146, 176], among others). In addition, different academic fields adopt very different citation practices (see [27] for a review on the topic), which results in a strong dependence of the mean number of citations on academic field, as shown in several works ([28, 124, 179], among others).

A natural question arises: how can we "fairly" use citation-based indicators to compare papers from different fields and of different age? The problem of comparing papers from different fields is usually referred to as the *field-normalization* problem. Several approaches to address this question have been proposed in the literature (see [247] for a recent review). A particularly simple approach is to divide each paper's citation count by the mean number of citations for papers of the same field published in the same year. The results by [179] suggested that this indicator, called *relative citation count*, produces a ranking that is statistically consistent with the hypothesis of a ranking that is not biased by field and age. This finding has been challenged by subsequent works by [6] and [249], which leaves the debate on age- and field- normalization procedures still open.

In this chapter, we analyze a large data set from Microsoft Academic Graph [213] to show that existing indicators of impact, including the relative citation count, fail to produce rankings that are not biased by age and field. To simultaneously assess these biases, we present a new procedure based on the Mahalanobis distance [127]. This permits to compare the ranking by a given indicator with those obtained with a simulated unbiased sampling, and hence

to quantify the overall ranking bias. An analytic result derived in Appendix allows us to assess the contribution of each field to the overall ranking bias. It is worth noticing that while we focus on the biases by age and field, our bias assessment procedure can be easily extended to detect any other kind of information bias.

We also present the first systematic study of the possible bias by field of the PageRank score [31] and of its age-rescaled version introduced by [130] at article-level. The motivation to analyze these network-based indicators comes from the finding that they outperform other indicators in identifying expert-selected milestone papers [130]. However, the application of PageRank and its variants to academic citation networks focused on data sets composed of papers from a single field [36, 129, 130, 244, 258, 260]. While the possible bias by scientific field of eigenvector-based algorithms has been explored by [248] at journal-level, the PageRank score's possible bias by academic field at article-level is still unexplored. We are the first ones to address it.

We introduce two novel indicators of impact motivated by the z-score: age- and field-rescaled citation count $R^{AF}(c)$ and age- and field-rescaled PageRank $R^{AF}(p)$. We find that the novel indicators produce paper rankings that are much less biased by age and field than the rankings produced by the other analyzed indicators. Nevertheless, also the Mahalanobis distance observed for the new indicators is not statistically consistent with ones obtained for a simulated unbiased process. This indicates that the problem of achieving an ideal unbiased ranking of the publications remains open.

The rest of this chapter is organized as follows: Section 2 summarizes the analyzed data set of publications obtained from the Microsoft Academic Graph. Section 3 reports bias by scientific field of four existing paper-level impact indicators. In Section 4, we introduce a rescaling procedure for citation count and PageRank scores motivated by the z-score. In Section 5, we introduce a general procedure to test for any kind of ranking bias, and present its application to assess the field and age bias of the rankings by the indicators studied here. In Section 6, we conclude by discussing possible limitations of our analysis and future research directions.

3.2 Data

We analyze a bibliographic data set which was provided for the KDD Cup 2016[2]. This data is a dump of the *Microsoft Academic Graph* (MAG) and contains more than 126 millions of publications and more than 467 millions citations [213]. Each publication is also endowed with various properties such as unique ID, publication date, title, journal ID, etc. We pre-processed the data (details are provided in 2.1.1A) to remove from the analysis papers with incomplete information, ending up with $N = 18\,193\,082$ unique publications and $E = 109\,719\,182$ citations.

The MAG has a field classification at paper level [213]. In the KDD cup dump, there are 19 main fields and numerous subfields up to 3 hierarchical levels of subsubfields. However, all the different subfields can belong to several main fields, meaning that each publication can belong to more than one main field. We use here the field classification at the highest hierarchical level, i.e., we only consider the 19 main fields. When calculating the citation count and PageRank score of papers (see Section 3.3), we consider the publications that belong to more than one field only once. In this way, we do not modify the number of citations that each paper receives and gives, and we do not change the topology of the network on which the PageRank scores are calculated. On the other hand, in agreement with [249], to compute the fields' size (see Tab. A.1 in Appendix A) and the field-rescaled indicators (see Section 3.3 and Section 3.4), each publication can be considered multiple times in the analysis, once for each field the publication belongs to. In this way, each field is represented by all its publications even if some of these are shared with other fields.

Before moving to the next Sections, we devote our attention to two main assumptions of our analysis. First, we assume that the Microsoft Academic data provide a representative sample of the population of publications and of their citations. This assumption is motivated by the findings of independent analyses of the Microsoft Academic data set [90, 98] that have shown that its

[2] `https://kddcup2016.azurewebsites.net/Data`

coverage is comparable to other popular academic databases, such as Scopus and Web of Science.

Second, to quantify the bias by field of impact indicators, we assume that the fields are given by the Microsoft Academic's field classification scheme at its highest hierarchical level. In the literature, there is no general agreement on which field classification scheme should be used to classify papers and there is an entire stream of works investigating issues related to this [2, 40, 178, 214, 261]. In particular, the choice of a suitable aggregation level has been shown to be delicate: by considering the most aggregate fields, heterogeneities in the subfields' citation patterns might be hidden [176, 236] – this effect has been shown to be magnified when iterative ranking algorithms are used instead of citation count [250]. On the other hand, increasing the resolution of the field classification may lead to largely-overlapping fields or to hardly interpretable fields. For example, [97] show that the MAG fields at the second highest level are too detailed and, for this reason, the authors suggest that they should not be used for field-normalization purposes. We leave to future research the important study of how different classification schemes impact the biases of rankings and how our results generalize to other data sets.

3.3 Shortcomings of existing indicators

In Sect. 2.3.1, have defined four existing indicators used to detect paper impact: citation count c, relative citation count c^f, PageRank score p, age-rescaled PageRank score $R^A(p)$. We now show that these indicators are severely biased by scientific field (Section 3.3.1).

3.3.1 Field bias of the existing indicators

After having described a set of existing indicators, we now apply them to the MAG data set to show that the rankings that they produce are biased by scientific field. For a ranking that is not biased by scientific field, the number of top-ranked publications from each field should be proportional to the

total number of publications from that field. In other words, for an unbiased ranking, we expect

$$\mu_f = \frac{z}{100} K_f \tag{3.1}$$

papers from field f among the top $z\%$ papers in the ranking, where K_f is the total number of publications from field f [179]. In the following, we denote by $k_f^{(m)}$ the number of publications from field f in the top-1% of the ranking by indicator m. We restrict our analysis to $z\% = 1\%$; results for other values of z are available upon request from the authors.

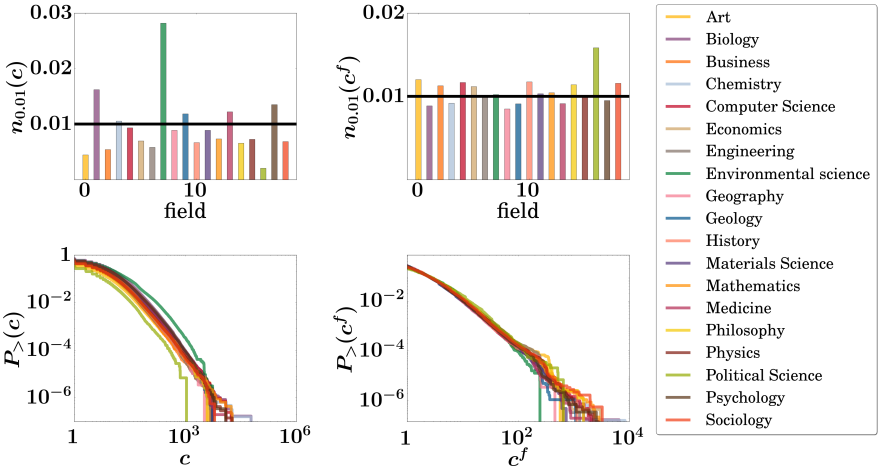


Figure 3.1: Field bias of the analyzed citation-based indicators. Top panels show histograms of the fraction of top-1% publications for each field in the ranking by (left to right) citation count and relative citation count. The black horizontal line is at 0.01, i.e., the expected value. Bottom panels show for each field the complementary cumulative distributions for citation count (left) and relative citation count (right).

In the top panels of Fig. 3.1, we illustrate the field bias of citation count, c, and relative citation count, c^f. The presence of strong biases is evident for both indicators because there are fields whose ratio $k_f^{(m)}/K_f$ is far away from the expected value 0.01. In particular, Environmental Science is extremely over-represented in the top of the ranking by citation count. We argue that this

bias comes from the fact that publications from this field have a mean citation count almost twice as big compared to publications belonging to other fields (see Table A.1). For relative citation count, we find a better agreement with what we would expect from an unbiased indicator. However, relatively large deviations are still evident, especially for the field of Political Science.

In the bottom panels of Fig. 3.1, we report the distributions of c and c^f for each field. These panels show that the bias by field is not limited to the top 1% papers in the ranking, but it arises from systematic differences between the score distributions across different fields. For example, when looking at the distribution of c, papers in the field of Political Science have consistently smaller probability to have more than one citation compared to other fields. For a detailed discussion about the bias of the ranking by c^f, we refer to Appendix B.

Fig. 3.2 reports the same analysis for PageRank scores, p, and age-rescaled PageRank scores, $R^A(p)$. This figure provides the first study of the dependence of PageRank score on academic field. The top panels of Fig. 3.2 show that the top positions of both rankings are biased by field, and both rankings overestimate the impact of publications in the field of Environmental Science. Again, we argue that this happens because the mean in-degree of publications from Environmental Science is approximately twice as big compared to publications that belong to other fields (see Table A.1). From the bottom-left panel of Fig. 3.2, we find that the full distribution of scores of Page Rank have a similar shape, but different broadness. These differences are slightly smaller for the age-rescaled PageRank, with the exception of the field of Environmental Science (see bottom right panel of Fig. 3.2).

3.4 New age- and field- normalized indicators

In this Section, we introduce two novel indicators of paper impact: the age- and field-rescaled citation count, $R^{AF}(c)$, and the age- and field-rescaled PageRank, $R^{AF}(p)$. The two indicators, $R^{AF}(c)$ and $R^{AF}(p)$, are obtained from citation count c and PageRank score p, respectively, through a rescaling procedure. This procedure is based on the z-score and is aimed at suppressing

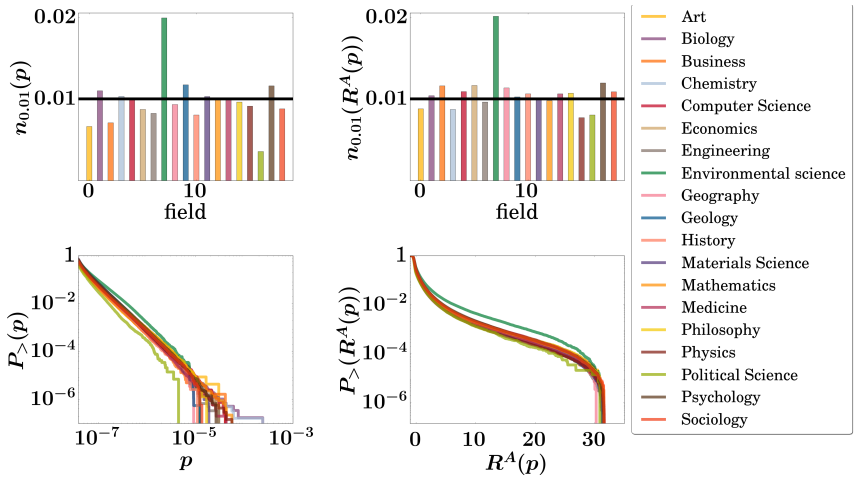


Figure 3.2: Field bias of the analyzed measures based on PageRank. Top panels show histograms of the fraction of top-1% publications for each field in the ranking by (left to right): PageRank and age-rescaled PageRank. The black horizontal line is at 0.01, i.e., the expected value. Bottom panels show for each field the complementary cumulative distributions for PageRank (left) and age-rescaled PageRank (right).

age and field bias. The idea of using the z-score is not new in scientometrics [28, 124, 130, 134, 146, 259]; our new indicators can be considered as variants of the indicator based on the z-score studied by [259] and their main difference is explained below.

3.4.1 Age- and field-rescaled citation count, $R^{AF}(c)$

To calculate the age- and field-rescaled citation count $R_i^{AF}(c)$ of a paper i belonging to a field f, we first compute the mean $\mu_i^{AF}(c)$ and the standard deviation $\sigma_i^{AF}(c)$ of the citation count of papers of the *same field* and of *similar age* as paper i. In particular, $\mu_i^{AF}(c)$ and $\sigma_i^{AF}(c)$ are computed over the papers that belong to the same field f as paper i and that are among the Δ_c closest

papers to i as measured by the distance $|i - j|$ between their rank by age. Then, the age- and field-rescaled citation count score $R_i^{AF}(c)$ is defined as

$$R_i^{AF}(c) = \frac{c_i - \mu_i^{AF}(c)}{\sigma_i^{AF}(c)}. \tag{3.2}$$

The averaging window size Δ_c is a parameter of the method, which we set to $\Delta_c = 1000$.

Differently from [259], for the computation of the z-score, we use temporal windows with the same number of publications, which in general corresponds to real-time intervals of different duration. This choice is supported by recent findings [159] that indicate that in citation networks, time is better defined by number of publications than by real time. Furthermore, rescaled indicators based on the z-score with fixed temporal-window duration have already been shown to under-perform with respect to the relative citation count c^f in the task of producing an unbiased ranking [259]. For these reasons, we do not include indicators based on z-score with fixed temporal-window duration in our analysis.

Differently from the relative citation count c^f, $R^{AF}(c)$ is expected to have not only uniform mean value across different publication dates and fields, but also uniform standard deviation. This should lead to a more balanced ranking of the papers. We show in the following that this is indeed the case.

3.4.2 Age- and field-rescaled PageRank, $R^{AF}(p)$

Previous works have shown that PageRank is biased towards old papers in scientific citation networks [36, 129, 133]. Moreover, we have shown in Section 3.3 that PageRank score p is biased by scientific domain. To simultaneously suppress these two biases, we propose the age- and field-rescaled PageRank score $R^{AF}(p)$. $R^{AF}(p)$ is defined similarly as $R^{AF}(c)$: we compute the mean value $\mu_i^{AF}(p)$ and the standard deviation $\sigma_i^{AF}(p)$ of the PageRank scores of the papers that belong to the same field as paper i and that are among the Δ_p closest

papers to i as measured by the distance $|i - j|$ between their rank by age. The age- and field-rescaled PageRank score is then defined as

$$R_i^{AF}(p) = \frac{p_i - \mu_i^{AF}(p)}{\sigma_i^{AF}(p)}. \tag{3.3}$$

In the following, we set $\Delta_p = 1000$.

3.4.3 Field bias of the new indicators

In the top panels of Fig. 3.3, we show that in the top-1% of the rankings by $R^{AF}(p)$ and $R^{AF}(c)$ each field appears well represented. In fact, the deviations from the expected value are very small especially if compared to the deviations of the other rankings (see top panels in Figs. 3.1 and 3.2). In the bottom panels of Fig. 3.3, we report that the the full score distributions for papers from different fields collapse well on top of each other thanks to the rescaling procedure.

3.5 Quantifying field and age biases

We begin this Section by introducing a new methodology to assess a ranking's bias based on the Mahalanobis distance (Subsection 3.5.1). Then, we use this to quantify the bias by field (Subsection 3.5.2) and the bias by age and field (Subsection 3.5.3).

3.5.1 A general framework to assess ranking biases based on the Mahalanobis distance

While Figs. 3.1 and 3.2 illustrate the substantial field bias of the existing indicators, the bias is much weaker (if any) for the new age- and field-rescaled indicators in Figure 3.3. Now we quantify this improvement by extending the statistical tests of bias suppression presented by [130, 177, 179, 249]. Similarly to these works, we assume that a ranking is unbiased if its properties are consistent with those of an unbiased selection process.

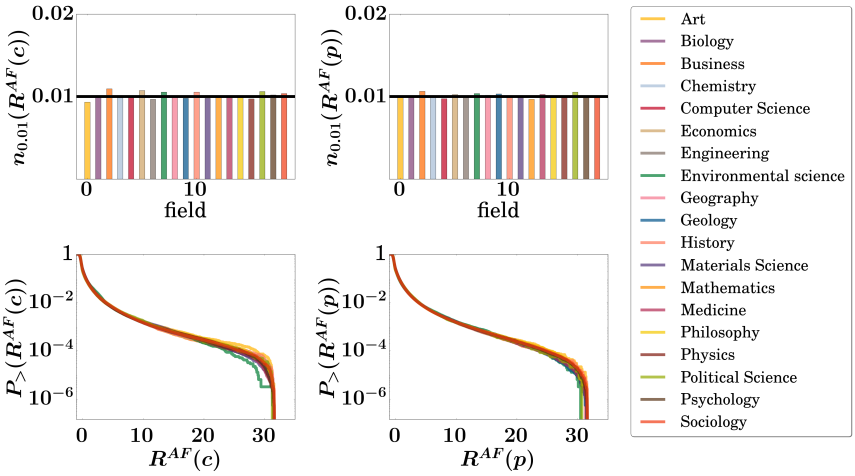


Figure 3.3: Field bias of our normalized indicators. Top panels show histograms of the number of top-1% publications for each field in the ranking by (left to right): age- and field-rescaled citation count, and age- and field-rescaled PageRank. The black horizontal line is at the expected value 0.01. Bottom panels show the complementary cumulative distributions for age- and field--rescaled citation count (left), and age- and field-rescaled PageRank (right).

Assessing the bias by field. Consider an urn which contains N marbles, each of them corresponding to one of the publications present in our data set. An *unbiased selection process* then corresponds to sampling from this urn at random without replacement a fixed number $n = \lfloor N \times 0.01 \rfloor$ of publications. From the extracted sample, we count the number of publications that belong to each field f, k_f, and record these numbers in the vector $\vec{k} = (k_1, ..., k_F)^T$; here F denotes the number of fields. The probability to observe a certain vector, $\vec{k}$, is given by the *multivariate hypergeomentric distribution* (MHD)

$$P\left(\vec{k}\right) = \frac{\prod_f^F \binom{K_f}{k_f}}{\binom{N}{n}} \tag{3.4}$$

where K_f is the total number of publications in field f. Following this selection process, among the n extracted publications, the expected number of publications for field f is $\mu_f = n\,K_f/N$.

Assume that the actual ranking by a given indicator m features $k_f^{(m)}$ publications from field f in the top 1% of its ranking. In general, the observed $k_f^{(m)}$ deviates from its expected value μ_f. A simple approach to quantify this deviation would consist in computing the z-score, defined as $z_f^{(m)} := (k_f^{(m)} - \mu_f)/\sigma_f$, where σ_f is the expected standard deviation for field f according to the MHD specified by Eq. (3.4). There are however two shortcomings of the z-score. First, the z-score only gives partial information for a MHD – how far from the expected values we are in units of standard deviations – but it does not provide information on how statistically significant the deviations are. Second, to quantify the overall bias of a given indicator m, we would need to aggregate the z-scores from the different fields. For example, we could take the average z-score, but this would neglect the correlation between the different fields coming from the constraint $n = \sum_f k_f^{(m)}$.

To overcome these two problems, we follow a different approach. We first run various numerical simulations that reproduce the unbiased selection process. These simulations produce a set of ranking vectors which are distributed according to Eq. (3.4) around the vector of expected values, $\vec{\mu} = (\mu_1, ..., \mu_m)$. Differently from [177], we do not estimate the confidence interval for the different fields separately. We calculate instead the Mahalanobis distance ($d_{\mathcal{M}}$, [127] and Appendix C) for each simulated vector from $\vec{\mu}$, and construct the distribution of $d_{\mathcal{M}}$'s obtained by the simulated unbiased selection process. The inset of the left panel of Fig. 3.4 reports the distribution of the $d_{\mathcal{M}}$ for 1 000 000 simulations. The distribution is centered around its mean value of 4.18 and the upper bound for the 95% confidence level is around 5.37. [3] For an *ideal unbiased ranking*, we would expect its $d_{\mathcal{M}}$ to fall into the 95% confidence interval of the distribution of the $d_{\mathcal{M}}$ obtained from the simulated unbiased sampling process.

[3] A curiosity for the reader. Here, the average of the square of the $d_{\mathcal{M}}$ for the unbiased sampling process is extremely close to the number of degrees of freedom of our problem. This stems from the fact that the MHD defined by Eq. (3.4) converges to a Multivariate Gaussian Distribution (MGD) as we increase the number of publications N while keeping n/N fixed and small. Our data set is large enough for this approximation to be accurate. The $d_{\mathcal{M}}^2$ of a MGD is distributed as a χ^2 variable with average equal to the number of degrees of freedom, i.e. 18 since we have 19 distinct fields and one constraint.

Assessing the bias by age and field The methodology presented above is easily generalized to simultaneously assess a ranking's bias by age and field. To add the temporal dimension to the bias assessment procedure, we split the publications into T equally-sized age groups, and repeat the above analysis by using $F \times T$ different categories of publications. In Section 3.5.3, we set $F = 19$ representing the number of fields and $T = 40$ as in [130], and thus we obtain 760 age-field groups of different sizes.

3.5.2 Results on the bias by field

The rankings produced by different indicators differ greatly by their $d_{\mathcal{M}}$ (see Fig. 3.4). As expected, the indicators that are not field-rescaled (c, $R^A(p)$, and p) are far from being unbiased. At the same time, relative citation count that is rescaled by field performs only slightly better than PageRank which is ignorant of any field information. The best results by a wide margin are achieved by our indicators, $R^{AF}(c)$ and $R^{AF}(p)$, obtained using the new rescaling. Nevertheless, both these indicators fail to meet the 95% upper bound achieved by simulated unbiased rankings. As disappointing as it may seem, this finding is not entirely surprising as the proposed rescaling procedures focus on equalizing the first two moments of the respective quantities (c and p) whereas the quantities' distributions can differ also by higher moments.

To understand which field contributes the most to the resulting $d_{\mathcal{M}}$ values, we have derived an alternative analytic expression for the $d_{\mathcal{M}}$

$$d_{\mathcal{M}}(\vec{k}, \vec{\mu})^2 = \sum_{i}^{F} z_i^2 \left(1 - \frac{k_i}{N}\right) \tag{3.5}$$

where we omit the indicator superscript (m) from the notation for z_i and $\vec{k}$ for simplicity. We have proven this formula analytically for $F = 3, 4, 5, 6$, and we have numerically tested it for $F = 19$ and 760 (see Appendix C); it remains open to prove it in arbitrary dimensions.

In Table 3.1, we report the individual fields' contributions to the $d^2_{\mathcal{M}}$ calculated using Eq. 3.5. We find that Biology and Computer Science are the fields which give the biggest contributions to the $d^2_{\mathcal{M}}$ for the rankings by citation count

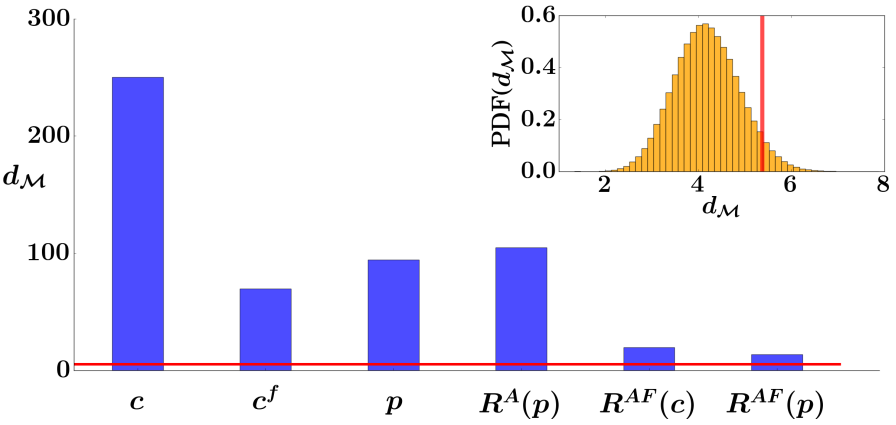


Figure 3.4: Mahalanobis distances, $d_{\mathcal{M}}$, for the analyzed indicators when considering the 19 main fields. From left to right: citation count, relative citation count, PageRank, age-rescaled PageRank, age- and field-rescaled citation count and age- and field-rescaled PageRank. The horizontal red line represents the upper bound of the 95% confidence interval obtained from the simulations. In the insets, we report the distribution of $d_{\mathcal{M}}$ coming from 1 000 000 simulations of the unbiased sampling process. Again, the red line represents the upper bound of the 95% confidence interval.

and relative citation count, respectively. This could not have been detected by looking at the deviations from the expected values. Indeed, in Fig. 3.1 we only see that Environmental Science and Political Science have the largest deviations. For the novel indicators, approximately one third of the $d^2_{\mathcal{M}}$ of $R^{AF}(c)$ is explained by the field of Economics and approximately one fourth of the $d^2_{\mathcal{M}}$ of $R^{AF}(p)$ is explained by the field of Mathematics.

In addition, we also find that the $d_{\mathcal{M}}$'s contributions across different fields assume values in a relatively broad range. This suggests that findings on rankings' bias by field may strongly depend on which disciplines are included or not in the analysis. We argue that arbitrary choices on which fields to include should be avoided in future research on field-normalization of impact indicators.

Field	c	c^f	p	$R^A(p)$	$R^{AF}(c)$	$R^{AF}(p)$
Art	1.15	1.95	3.01	0.31	2.84	0.09
Biology	**36.46**	15.81	6.74	0.87	0.12	0.16
Business	2.06	2.08	5.95	1.51	14.56	13.50
Chemistry	0.34	8.44	0.85	9.28	4.23	0.00
Computer Sc.	0.29	**23.86**	0.02	3.09	0.12	13.50
Economics	3.34	6.51	4.20	5.77	**32.32**	7.93
Engineering	8.36	0.09	10.48	0.35	8.36	0.03
Environmental Sc.	16.82	0.04	**35.67**	29.89	2.45	2.23
Geography	0.05	1.34	0.15	0.50	0.15	0.51
Geology	1.02	3.04	6.42	0.13	2.10	10.06
History	0.69	2.48	1.67	0.15	3.12	0.52
Material Sc.	0.39	0.43	0.23	0.01	2.37	0.10
Mathematics	5.17	1.94	0.10	0.14	0.09	**25.34**
Medicine	4.02	7.66	0.06	1.94	2.16	19.56
Philosophy	1.49	3.23	0.12	0.36	0.15	0.07
Physics	8.30	0.21	6.00	**33.67**	14.34	0.01
Political Sc.	1.47	10.22	6.82	0.51	1.47	2.44
Psychology	5.75	1.43	8.56	10.24	2.93	1.65
Sociology	2.83	9.23	2.95	1.30	6.11	2.30

Table 3.1: The individual contribution $z_i^2\,(1 - k_i/N)$ of each field i to the $d_{\mathcal{M}}$ by the different indicators.

To summarize, our bias suppression test allows us not only to estimate the level of bias ($d_{\mathcal{M}}$) of the various indicators, but also to quantify which percentage of the total bias ($d^2_{\mathcal{M}}$) of an indicator is explained by each single field.

3.5.3 Results on the bias by age and field

While the analysis of the previous Subsection focused on the ranking bias by field, in this Subsection we use the $d_{\mathcal{M}}$ to simultaneously assess the bias by age and field of a given ranking.

In Fig. 3.5, we show the $d_{\mathcal{M}}$'s for the different indicators and for the 95% confidence interval for the simulated unbiased selection process using 40×19 age-field types of publications. For citation count, PageRank, relative citation count and age-rescaled PageRank we have to reject the hypothesis that the

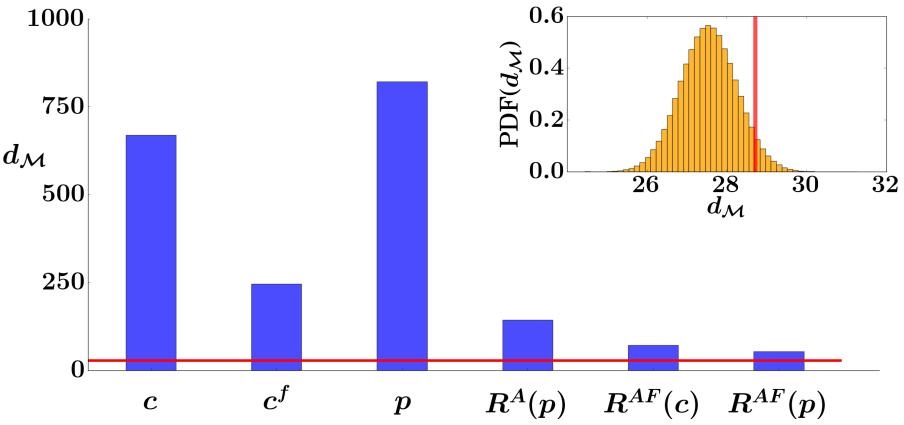


Figure 3.5: Mahalanobis distances, $d_{\mathcal{M}}$, for the analyzed indicators when considering the 760 age-field groups. From left to right: citation count, relative citation count, PageRank, age-rescaled PageRank, age- and field-rescaled citation count and age- and field-rescaled PageRank. The horizontal red line represents the upper bound of the 95% confidence interval obtained from the simulations. In the insets, we report the distribution of $d_{\mathcal{M}}$ coming from 1 000 000 simulations of the unbiased sampling process. Again, the red line represents the upper bound of the 95% confidence interval.

rankings of these indicators are not biased by age and field. For the improved indicators, age- and field-rescaled citation count and PageRank, we also have to reject the null hypothesis, even though they are much closer to the 95% confidence interval.

3.5.4 Simultaneously visualizing the bias by age and field

To visualize the field and age bias of the rankings by the analyzed indicators, we use heat maps in the age-field group plane (see Fig. 3.6). In these heat maps, each cell represents a field-age group, and its color indicates the level of bias. A white cell indicates that the number of papers in the respective age-field group falls into the 95% confidence level ($\mathcal{C}_{95\%}$) determined with the simulations. Hence, white means that no bias is detected for that age-field group. While for representing the bias towards or against a group of papers,

we use blue (overestimation) and red (underestimation). To obtain a range of over/under-estimation, the brightness of the colors ranges from white (no bias) to intense blue/red. The most intense colors indicate that the number of papers from that age-field group is 5 standard deviation smaller/bigger than the expected value.

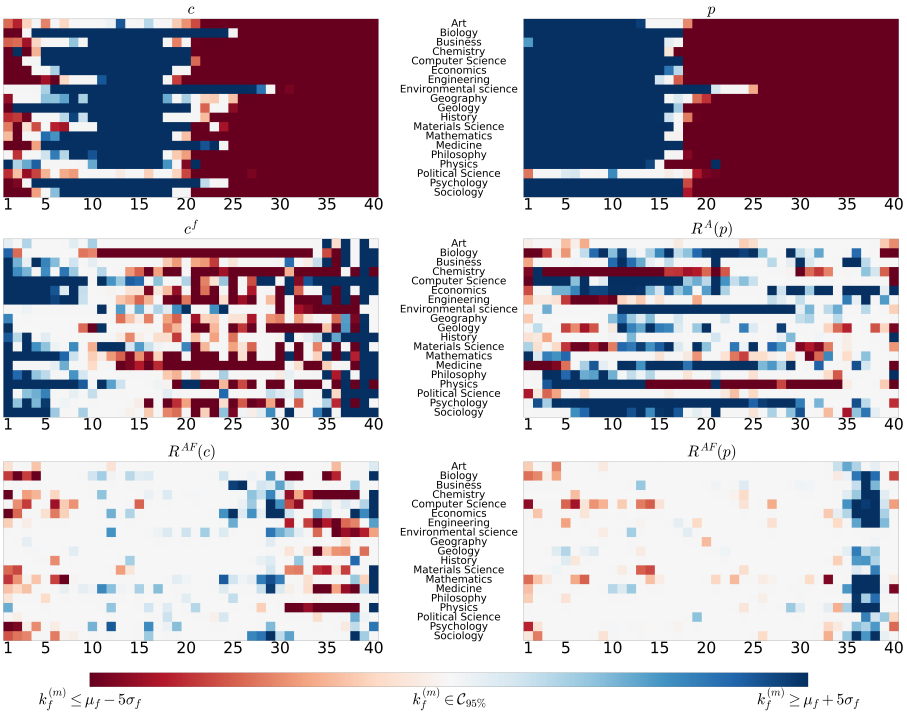


Figure 3.6: Heat maps showing the bias by field and age of the rankings by the different indicators. Each cell represents an age-field group: age groups are represented horizontally, while fields are represented vertically. The color of the cells shows the bias of the indicators with respect to that age-field group. White means that the respective age-field group is fairly represented in the top 1% of the ranking by the indicator. While we use a color scale from white to intense red (blue) for age-field group which are underestimated (overestimated).

The top panel of Fig. 3.6 shows that, independently of field, citation count and PageRank systematically over-represent old papers and under-represent recent papers. This is in agreement with the findings of several other works [36, 129, 130, 146]. The only exception is Political Science which is usually underestimated independently of paper age. We argue that this happens because this is the smallest field in the data set, and it has become an academic discipline by itself much later compared to most of the other fields [4]. Also, the oldest papers in most fields are under-represented by citation count, which reflects the change of citation practices over time.

The middle panels of Fig. 3.6 show that the relative citation count and age-rescaled PageRank suppress large part of the biases of the original indicators, yet specific fields are consistently overestimated or underestimated. For example, both age-rescaled PageRank and relative citation count under-represent papers belonging to the field of Chemistry. A peculiarity of relative citation count is that it over-represents both the oldest as well as the most recent papers at the cost of the other papers.

The bottom panels of Fig. 3.6 show the heat maps for the new indicators age- and field- rescaled citation count $R^{AF}(c)$ and PageRank $R^{AF}(p)$. We find that the respective rankings are much less biased towards specific fields compared to all the other analyzed measures. However, there are two patterns: for $R^{AF}(c)$ recent publications tend to be underestimated for some fields, whereas for $R^{AF}(p)$ recent publication tend to be overestimated for almost all fields. These rather systematic patterns must have their roots in changes of the citation and PageRank score distributions with time. Since our rescaling procedure was fixing the first two moments of these distributions, the observed patterns come from differences in higher moments. Thus, the distributions of $R^{AF}(c)$ and $R^{AF}(p)$ are aligned only partially for papers of different age.

[4] We notice that the classification of Political Science as one of the highest-level fields is not obvious. In Scopus categories, "Political Science and International Relations" is only a subfield of the higher-level field Social Science [`http://www.scimagojr.com/journalrank.php?area=3300`]. In the Web of Science classification scheme, "Political Science" is only a subfield of the higher-level field "Social Sciences, General" [`http://ipscience-help.thomsonreuters.com/inCites2Live/8300-TRS.html`].

3.6 Conclusion

To summarize, in this chapter we have analyzed a large citation network from the Microsoft Academic Graph to show that the rankings of papers by well-known indicators are extremely biased by age and field. The level of bias of the rankings has been quantified with a new statistical framework based on the Mahalanobis distance. This framework has allowed us to simultaneously quantify the age and field biases of the analyzed rankings, and to determine which groups of papers give the largest contributions to the observed bias. To allow other researcher to easily implement our statistical test for ranking bias, we make the respective code publicly available[5] together with a quick tutorial on how to use it[6]. In addition, we have also introduced two new indicators of paper impact, rescaled citation count $R^{AF}(c)$ and rescaled PageRank $R^{AF}(p)$ that produce much less biased rankings than existing indicators. In particular, the ranking by $R^{AF}(p)$ is approximately three times less biased compared to the least biased existing indicators, relative citation count and age-rescaled PageRank.

The contribution of our results to the debate on the validity of field normalization procedures is threefold. First, our findings are in agreement with the conclusions of [6] and [249] which argued that the relative citation count introduced by [179] can be insufficient to effectively remove citation count's bias by age and field. Second, we show the importance of testing indicators using an accurate statistical procedure, such the one introduced here. Indeed, for the least-biased indicators analyzed, $R^{AF}(c)$ and $R^{AF}(p)$, no clear indication of bias is found at first glance. However, when using the statistical test based on the Mahalanobis distance, we find a significant discrepancy between their rankings and those coming from unbiased sampling process. We argue that including higher-order momenta (such as the skewness) in the rescaling procedure can be an efficient way to further reduce the rankings' level of bias. Third, by deriving an explicit formula to calculate the contribution of each field to the bias of a ranking, we find that the these contributions assume a broad range of values. We obtain similar findings also for the contributions

[5] `https://github.com/giava90/quantifying-ranking-bias`.
[6] `https://www.sg.ethz.ch/team/people/gvaccario/quantifying-ranking-bias/`.

to the age-field bias. This means that the level of bias of rankings depends heavily on which years or fields are included in the analysis. For this reason, in future research on age and field normalization of indicators, it is essential to clearly motivate which years and fields are included in the analysis, avoiding arbitrary or uncritical decisions.

To address the bias by age and field of ranking of papers, we have first divided the papers in groups with similar age and from the same field. Then, we considered only the sizes of these groups to define an unbiased selection process from which we obtained a statistical null model for an unbiased ranking. In principle, additional information can be included into the null model to correct for other effects. For example, including information about the co-authorship network would permit to correct for the effect of this network on the growth and structure of the citation network [193, 194]. In this way, we would gain a better understanding of how the social dimension of science contributes to the field and age biases of impact indicators.

We emphasize that removing the biases addressed in this work and those that come from social aspects is of primary importance not only for scholarly publication databases, but also for several other information systems, such as the WWW or online social networks [202]. As a matter of fact, every day scholars and on-line users explore available knowledge using recommender systems based on ranking algorithms. This challenges us to design more sophisticated filtering and ranking procedures to avoid biases that can systematically hide relevant contents or only show information too similar to what the users already know.

To conclude, by reducing the age and field biases from indicators of scientific impact and by extending the existing statistical tests for biases, we contribute to the challenge of quantifying and suppressing biases of rankings in information systems.

Chapter 4

The empirical flow of knowledge at the journal level

Summary

We investigate the importance of journals by using a large citation data set from Microsoft Academic Graph (MAG). For our investigation, we adopt a *path* perspective to reconstruct the knowledge flow among journals from citation data. We show that this is radically different compared to the *network* perspective often used in citation analysis. Indeed, with the *path* perspective, we retain the empirical flow of knowledge and ideas that are generally discarded by the *network* perspective. Based on this approach, we propose new indicators to determine the similarity and influence of journals. Finally, we compare our indicators with established ones, such as Bibliographic coupling and PageRank, computed using the *network* perspective. From this comparison, we validate our approach based on the *path* perspective and un-hidden new patterns present in the data. [1]

[1] Based on [234]

4.1 Introduction

In academia, journals have mainly two roles: the evaluation of new works and serving as basis for academic credit. Especially for this second role, scientometricians and bibliometricians have been developing indicators that capture the importance of journals. With the increasing availability of citation data, more and more indicators are developed using citation analysis often in combination with a network perspective. However, indicators based on this perspective need *a path transitivity assumption* that is not justified for the citation network at the journal level. To overcome this problem, we develop new reliable indicators of journal importance by combining two modeling approaches based on *path abstraction* and *higher-order networks*. We show that our new indicators are more reliable compared to established ones as they better capture the empirical flow of knowledge occurring between journals.

In order to develop quantitative indicators, scientometricians and bibliometricians traditionally use citation analysis. This analysis consists of identifying properties of documents trough their cross-referencing. One example is the commonly used impact factor [66, 67]. It captures the influence of journals by computing the average number of citations received by papers belonging to them. More sophisticated indicators have been obtained by combining citation analysis with a network perspective. Researchers have adopted this perspective by constructing a citation network at the journal level. In this network, nodes are journals and links are citations among papers published in the journals. Network measures, such as eigenvector and betweenness centralities, have been proposed as indicators to determine the influence [164] and interdisciplinarity [119] of journals.

The use of network measures on citation data lies on the assumption that knowledge flows in the opposite direction compared to citation links. This means that a paper, i.e., a knowledge artifact, receiving many citations contains knowledge that is often re-used to create new knowledge, i.e., new papers. At the journal the level a similar statement holds, namely that citation links among journals capture how knowledge flows from one journal to another. Additionally, most network measures depend also on the *path transitivity as-*

sumption: when inferring (from data) the existence of links from A to B and from B to C, we automatically permit the existence of a path of length two from A to C via B. For example, this assumption is necessary to construct paths from citation links at the journal level. These paths represent flows of knowledge between journals and hence, they have been used to compute journals' similarity [216], influence [164] and interdisciplinarity [119].

However, the path transitivity assumption is not justified in the citation network at the journal level for two reasons. The *first* reason depends on the *projection* of the citation links from the paper to the journal level. This procedure changes the single units of analysis from papers to journals, and this change invalidates the transitivity assumption. Indeed, given two consecutive links in the citation network $\{(A, B), (B, C)\}$, we do not know if the paper in B cited by the paper in A is also the paper citing the paper C. This means that we cannot know if there was any knowledge exchanged by A or C via B. With the path transitivity assumption, we would instead assume the presence of a path between A and C via B. In other words, it would assume a fictitious knowledge mixing between papers in B and overestimate the flow of knowledge between journals.

The *second* reason that invalidate the transitivity assumption comes from the *time aggregation* of citation links. When we aggregate paper citations information into journals, we aggregate citations belonging to papers published at different times. This procedure discards timing information of the citations and does not preserve their empirical temporal ordering. This implies that two consecutive citation links on the citation network at the journal level (e.g., $\{(A, B), (B, C)\}$) could have appeared in the same temporal order in the data (i.e., first (A, B) and then (B, C)). Thus, the knowledge went first from B to A and then from C to B. By this, the two journals A and C cannot share any knowledge using the middle journal B. While, by assuming path transitivity, one would infer the existence of a knowledge flow that was never observed in the data. Again with this assumption, one overestimates the knowledge exchanged between journals.

To overcome the problems introduced by the path transitivity assumption, we use *path abstraction*, meaning that instead of concentrating on single citation

links, we look at consecutive citations between papers. These citations represent the knowledge flow at the paper level, and we can use them to construct sequences of papers, i.e., paths on the citation network at the paper level. Then, we project such paths at the journal level and obtain a set of journal sequences that match the empirical flow of knowledge. With these paths at the journal level, we can investigate the role of journals in the dynamic process of knowledge diffusion. Additionally, by using the statistical test developed in [201], we recover a network perspective. With this test, we detect the more statically significant paths in the empirical data and how to represent them as links and nodes in a higher-order network [204]. Following [201], we name such a network the optimal-order network, and on this network, we apply network analytic methods.

In order to quantify the empirical flow of knowledge, this chapter focuses on one centrality measure and one algorithm: PageRank [31] and `Infomap` [188]. The former is used to rank nodes, while the latter to cluster them (see Chap. 2.3.1 and Chap. 2.2.3 for their definition). Both of them are defined using a diffusion process on the network, and hence, both of them capture different aspects of the knowledge flow among journals. In particular, with PageRank, we rank journals depending on how important they are in the knowledge diffusion process. Whereas with `Infomap`, we group journals depending on how similar is the knowledge traversing them. We are not first one to rank journals using PageRank [164] or to cluster them with `Infomap` [188, 190]. However, we are the first one to apply this measure and algorithm on the optimal-order network detected from the empirical paths.

The remaining of this chapter is divided in the following way: Sect. 4.2 clarifies the pitfalls in analyzing citations data with a network perspective at the journal level and how they can be avoided using path abstraction. Then, we apply the standard network approach and the path abstraction to analyze citation data coming from the MAG in Sect. 4.3. We show that the reconstructed knowledge flows are different and how this influences the ranking of journals according to PageRank. In Sect. 4.4, we apply `Infomap` on the optimal-order network and define new measures to classify journals and compute their similarity. Also, we validate our new measures and compare them against established ones. We conclude and summarize our results in Sect. 4.5.

4.2 The network and the path perspectives

Before jumping into the definition of new indicators to determine journals influence and similarities, let us discuss why we have to use *path* abstraction to analyse citation data at the journal level. In citation data, we usually have a set of documents $D = \{p_1, p_2, .., p_N\}$, and a set of citations among them $C = \{(p_2, p_1), (...), ...\}$ where (p_j, p_i) represents a citation from document p_j to p_i with $i < j$. For the sake of notation, we assume that the subscript of documents represents an ordering by age: so for example, p_1 is older than p_2 and p_2 is older of p_3 and so on and so forth.

4.2.1 Why the network perspective fails at the journal level

We restrict our attention to the case where the documents in D are scientific *papers* published in *journals*[2]. From the sets of papers, D, and of citations, C, we can reconstruct a citation network at the *paper* level, where nodes are the papers in D and links are the citations in C. One could argue that to investigate the citation network at the *journal* level we could define a new network where (a) nodes are journals that contain the papers in D, and (b) links are the citations in D projected at the journal level. Even thought the first part is correct, the second part will make us lose information about how knowledge flows between journals. To understand why this is the case, consider the following situations:

1. we have 4 papers $D = \{p_1, p_2, p_3, p_4\}$ and three journals $J = \{A, B, C\}$. The younger paper, p_4, belongs to journal-A, the second and third papers, p_2 and p_3, belong to journal-B and the older paper, p_1, belongs to journal-C. Additionally, we have the following citations $C = \{(p_4, p_3), (p_3, p_1)\}$ (see Fig.4.1(a)).

[2]The problem that we are about to address goes beyond the specific data that we are here considering. As we will see, the failure of a network perspective lies in a projection needed to aggregate information, and it is not dependent on particular details of the bibliographic data used.

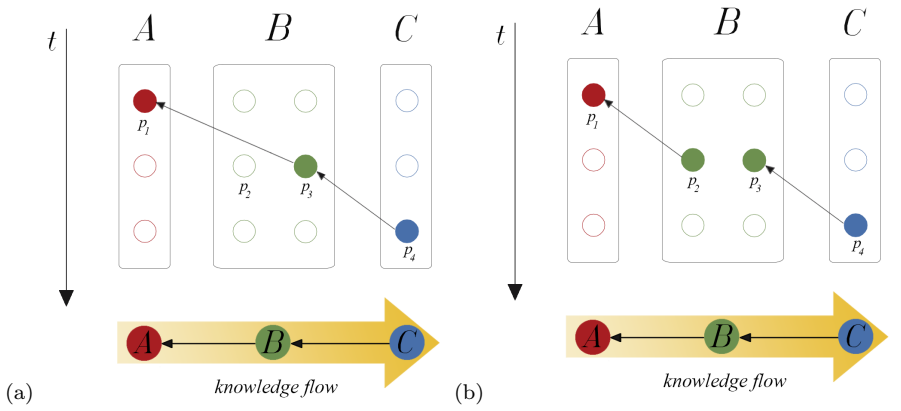


Figure 4.1: The citation projection from the paper to the journal level. In (a) we illustrate the case where citation links allow for knowledge to flow from journal A to journal C via journal B. This knowledge flow is correctly captured at the citation network at the journal level. While in (b), we have the case where citation links *do not* allow knowledge to flow from A to C via B and this is *not captured* at the citation network at journal level.

2. we have the exact same setting as before, but we change one link: instead of (p_3, p_1), we have (p_2, p_1), i.e. $C' = \{(p_4, p_3), (p_2, p_1)\}$ (see Fig.4.1(b)).

In Fig. 4.1, we reconstruct the citation network at the *journal* level for both situations by aggregating and projecting the citations from the papers to the journals. Here, we find that the citation networks at the *journal* level are the same, i.e. for both situations we obtain the network $G(V, E)$ where $V = J = \{A, B, C\}$ and $E = \{(A, B), (B, C)\}$. However, we had two different citation network at the *paper* level as $C \neq C'$. What do we miss by looking at the citation network at the *journal* level? Unfortunately, precisely what matters to study the flow of knowledge. In the first case, Fig. 4.1(a), we see that knowledge and information can be propagated from journal-C to journal-A via journal-B. While in the second case, see Fig. 4.1(b), this cannot happen. However, when looking at the citation network at the journal level, we cannot detect such a difference. In other words, by adopting a (standard) network perspective, one wrongly infers knowledge flows between journals.

Citation analysis and a network perspective at the journal level do not work as we project too naively information from the *paper* level to the *journal* level. The projection at the *journal* level introduces a fake mixing of knowledge between papers belonging to the same journals. In the above example, when we project we are indirectly assuming that there is a flow of knowledge between p_2 and p_3 just because they belong to the same journal (see Fig. 4.1). However, such internal knowledge mixing does not exist. In the next section, we show how to solve this problem.

4.2.2 Recovering the empirical flow of knowledge at journal level

To solve the fake mixing of knowledge, we use the concept of time respecting paths developed in [190, 201, 257]. In these works, the authors propose a framework for modeling sequential data, such as click streams and travel patterns, as paths on networks. In sequential data, a time respecting path is a sequence that respects the observed temporal ordering (e.g., sequences of web-pages visited by a user or sequences of airports visited by a traveler). We adopt this approach to analyze citation data as also this data can be seen as sequential data: each outgoing citation can be interpreted as a time-stamped interaction between the citing and cited papers. By following the citations among the different papers, we construct paths on the citation network at the *paper* level. In oder words, we follow citations from the citing to the cited papers and obtain sequences of papers that respect the time ordering of the citations.

The citation network at the paper level is a *Directed Acyclic Graphs (DAG)*. In this type of network, if we follow directed links from a starting node, it does not exist a path bringing us back to this starting node. This happens as an older paper cannot cite a younger one. Each cited paper/node in the network can be seen as the origin (roots) of some paths, and we can follow its citation links (in the opposite direction) towards younger papers (leaves). By connecting all the root to the leaf papers, we can reconstruct all the paths via which knowledge has been recombined and propagated. These paths represent the empirical flow of knowledge.

Once we have reconstructed the empirical flow of knowledge using sequences of papers, we project such sequences at the journal level. This means that we replace the papers inside the sequences with the journals to which the papers belong. By this, we obtain sequences of journals that correctly represent the empirical citation patterns. Now that we have reconstructed the empirical knowledge flow at the journal level, how can we use this to analyze journals with network analytics methods? We answer this question in the next section.

4.2.3 Recovering a network perspective

We use *higher-order network* models to represent journal sequences with a network perspective. These models have been developed in [203, 204, 257] and extended to path data in [201]. To understand what higher-order networks are, we can consider a first-order network as the "standard" aggregated network, e.g., where nodes represent journals and links citations. Then, a second-order network is a network where nodes are links in the "standard" network and links are paths of length two observed in the data, e.g.. two consecutive citations among three papers, (p_k, p_j) and (p_j, p_i) with $i < j < k$. Similarly, a third-order network is a network where the nodes are the links in the second-order network, and the links are paths of length three observed in the data. In general, on the K-order network, nodes represent observed paths of length-K, and links represent observed paths of length-$(K + 1)$.

Our data contains paths of different lengths. When a data set contains paths with different lengths, we can use a multi-order graphical model [201]. A MOG combines higher-order network models from order $K = 0$ to an optimal maximum order K_{opt}. If the data contains paths with a maximum length of K_{max}, then one could preserve all the details of the data by choosing $K_{opt} = K_{max}$. However, this causes overfitting, i.e., the multi-order model obtained would model only the analyzed data. Instead, a good model should capture only patterns that can be generalized also to other unseen data. Using a statistical test, the author of [201] provides criteria to balance between "keeping all the details of the data" and "overfitting". The test is based on a Maximum Likelihood approach and it allows us to determine the optimal maximum order, K_{opt}.

In the next section, we will apply the statistical test of [201] to our citations pathways to compute the K_{opt}. Then, we will represent our paths using a K_{opt}-order network instead of the full multi-order graphical model. The development of measures that analyze the multi-order graphical is still ongoing. Whereas, it has been shown that the *optimal*-order network well models both synthetic and empirical data [201].

Why does it work? When choosing the optimal order, we are again projecting paths into a (higher-order) network. Hence, we are introducing *again* a fictitious mixing of knowledge among papers belonging to the same journals. However, now the mixing mainly occurs between those sub-sets of papers that share the same incoming and outgoing links at the journal level. This happens as the optimal-order network encodes in its topology those paths more frequently observed.

We visualize and summarize the discussed procedure in Fig. 4.2. The citation data at the paper level allow us to create a DAG Fig. 4.2(a). From a DAG, we can construct a citation network at journal level in two ways: by projecting the citation Fig. 4.2(b), or by projecting the paths Fig. 4.2(c). The former method destroys the empirical knowledge flow between the journals, while the latter preserves it.

We use the above presented methodology in the next section in order to analyze real citation data coming from the `MAG17` . By this, we represent the `MAG17` citation data as a higher-order network. After reconstructing this network, we apply network-analytic tools to investigate the role of journals in the empirical flow of knowledge.

4.3 Reconstructing the knowledge flow

We use information coming from the Microsoft Academic Graph (`MAG17`). In particular, we use a dump of the data-set coming from OpenAcademic web-portal[3] released on the 2017-06-09. These are nine zipped files of about 13 GB

[3]https://www.openacademic.ai/oag/

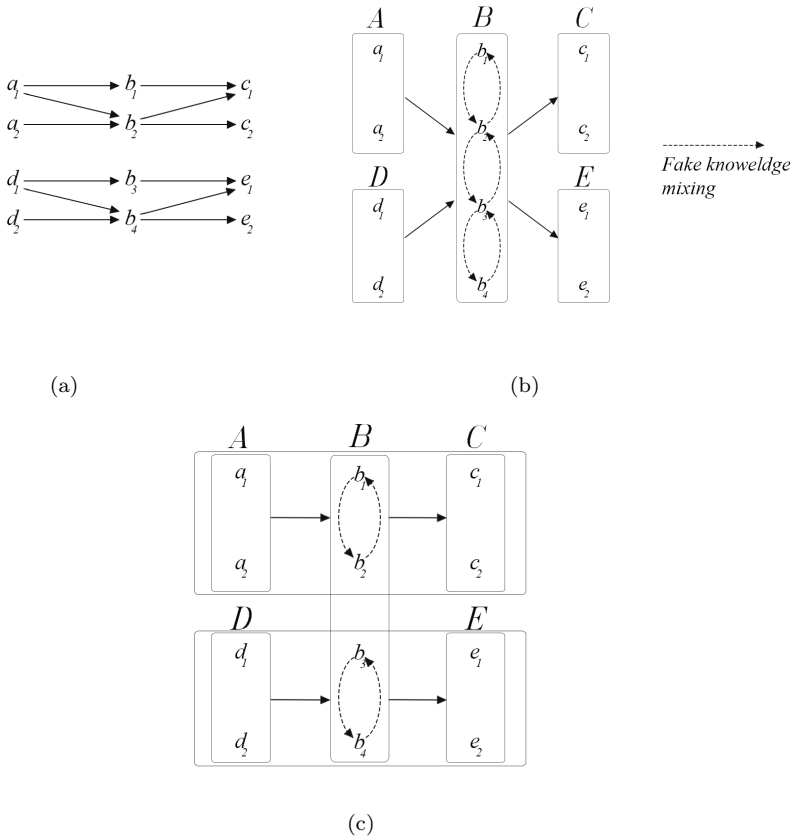


Figure 4.2: Summary of the citation projection from a *DAG*. Given a set of publications $\{a_{1,2} \in A, b_{1,2,3,4} \in B, c_{1,2} \in C, d_{1,2} \in D, e_{1,2} \in E\}$ and citations, we can construct a *DAG* (a). Then, we can project the citations at the journal level and obtain (b) where we observe a fictitious knowledge mixing among all the papers belonging to B. Otherwise, we can project the paths and obtain (c) where the fictitious knowledge mixing occurs only among smaller sets of papers sharing the incoming and outgoing citations.

each. The data contains 166 192 182 papers with different information, such as the DOI, venues in which was published and the reference list [213, 220].

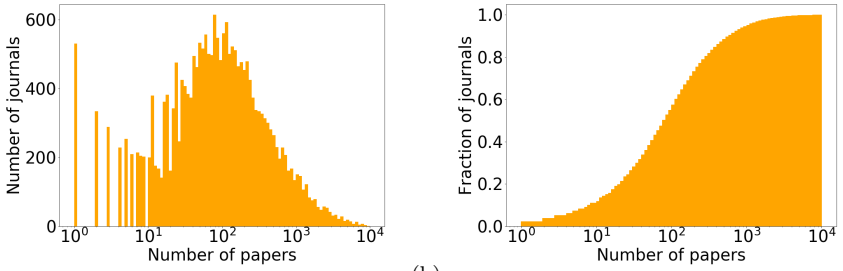


Figure 4.3: Number of papers per journal. (a) Histogram of the number of journals with a given number of papers. (b) Percentage of journals with at least a given number of papers.

papers	citations	journals	links (unique)	paths	[Min, Max]
109 969	150 000	94	3 038	722 999	[1, 22]

Table 4.1: Key statistics of the data.

The `MAG17` is a massive data-set, and to explore all of it is computationally challenging. When using only one file, we find that there are more than 20 000 journals and about 8 million papers out of 45 million have a DOI. In addition, about 2 500 of these journals appear less than ten times in our data. This means that there are 2 500 journals that contains less than ten papers. However, more than 80% of the journals have at least 100 papers (see Fig. 4.3).

Among the 20 000 journals, we concentrate on 100 main ones according to Google metrics [4]. We concentrate on top-journals as for these journals we have higher statistics, and they tend to be more known. This will allow us to obtain more statistically significant results and more interpretable results. Note that deciding the top journals according to Google metrics is not particularly limiting. Most rankings contain in their top positions the same journals, only their relative rankings usually change. Also, the top journals according to Google metrics belong to various scientific disciplines. Hence, our results will not be restricted to citation data coming from one scientific discipline.

[4] `https://scholar.google.ch/citations?view_op=top_venues` retrieved on the 01/06/2018.

In order to use the journal list coming from Google, we have to match the names from the `MAG` data to the names provided by Google. From the 100 journal names coming from Google metrics, we were able to match 94 journals (see Appendix D for more details). For the 94 matched journals, we project both the `MAG` citations among papers and the paths reconstructed from these citations. Both these steps have been accomplished using `Pathpy`, an open source python package available on `GitHub` [5]. To our knowledge, this is the only package that allows for such type of analysis and hence, it provides the unique opportunity to perform the following analysis. By using such a package, we reconstruct the DAG from the raw citation data, extract the paths at the paper level, and project them to the journal level. Note that the path extraction process is computationally costly with increasing citation links. For the analysis in this chapter, we only considered 150 000 citation links among 94 journals. This clearly limits our results to the sub-sample of citations analyzed.

In Table 4.1, we report key statistics of the paths extracted from the `MAG17`. From 109 969 papers with 150 000 citation links between them, we obtain 232 182 unique paths out of 722 999. The 722 999 paths generate 24 496 391 sub-paths between our 94 journals. Note how from $\sim 10^6$ citation links, the number of paths reconstructed paths is $\sim 10^8$, i.e., it is two order of magnitude bigger! Also, note that the citation network at the journal level is much denser compared to the citation network at the paper level. This result is expected as we are projecting and aggregating 150 000 links between 109 969 papers to 94 journals. At the same time, we believe that the number of observed unique links at the journal level (3 038) is small. Indeed, we argue that we would expect a network with more different links if we were randomly projecting citations from the paper level to the journal level. To verify this statement, we would need to define a null for the randomized projection, and this goes beyond the scope of the present chapter. We leave such a study for the future.

The longest paths are three and have length 22. They connect Cell Stem Cell, Nucleic Acids Research, and Nature Communications to Gut. By checking the publication history of these journals, we find that Cell Stem Cell, Nucleic Acids Research, and Nature Communications are all at least ten years younger than

[5] `https://github.com/uzhdag/pathpy`

Gut. This age difference partially explains why we find Gut at the end of the longest paths. Indeed, in our sample, we also have older journals like Nature that was created in 1869, i.e., 90 years before Gut. However, Nature is not at the end of the longest paths. We believe that the reason explaining why Gut is at the end of the longest paths goes beyond its age and we leave this for future analysis[6]. The most frequent path is of length one, and it connects Monthly Notices of the Royal Astronomical Society with itself. This path occurs 614 318 times when also considering its appearance as a sub-path. The most frequent path connecting two different journals connects Cell to Science, and it occurs 108 539 times (when also considering its appearance as a sub-path).

4.3.1 Difference between projecting citation links and paths at the journal level

To compare the standard network approach and our path abstraction, we study the difference between the first-order network and the higher-order network constructed using citation paths. Recall that the first-order network is a network where nodes are journals and links are citations between papers belonging to the journals. A second-order network contains pairs of journals as nodes and links as citation paths at the paper level of length 2.

Among the various possible higher-order network, we analyze the optimal-order network as defined in [201]. Using the statistical test presented by Scholtes [201], we find two as optimal order. This means that when a paper in a journal-A cites a paper in a journal-B, the out-going citations of this last paper in B strongly depends on journal-A. This is a first interesting result. Indeed, even though one would argue that the out-going citations of a paper depend on its incoming citations, it is not trivial to find that this dependency extends at the journal level.

To study the differences between the first-order and the optimal-order network, we focus on measures defined using diffusion processes. These processes are based on random walk processes and are relatively simple. However, they are extremely good at capturing nodes and network properties. Indeed, diffusion

[6]It might even be an artifact of the citation sampled for this study.

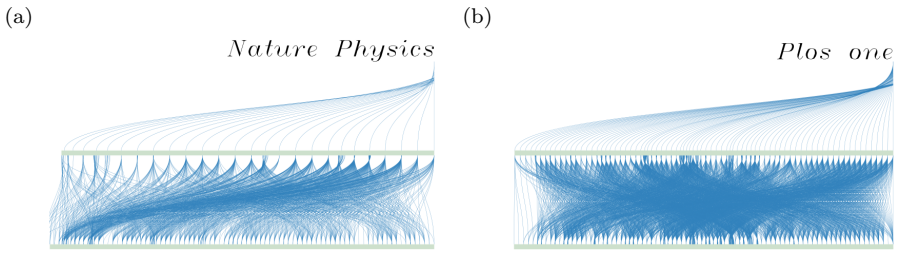


Figure 4.4: Alluvial diagrams on the first order network for Nature Physics (a) and Plos One (b).

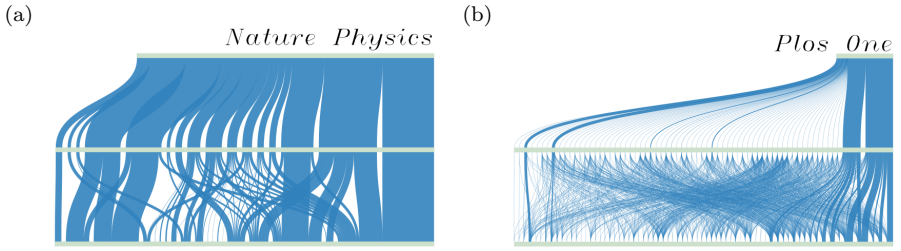


Figure 4.5: Alluvial diagrams on the optimal order network for Nature physics (a) and Plos One (b).

processes are at the base of successful network centrality measures, such as PageRank, and clustering algorithm, such as `Infomap`. For more details about PageRank and Infomap see Sect. 2.3.1 and Sect. 2.2.3.

We use the alluvial diagrams introduced in [111, 189, 190] to study diffusion processes on the first and the optimal order network. In Fig. 4.4, we visualize a diffusion process evolving on the *first*-order network from two journals (Nature Physics (a) and Plos One (b)). The diffusion process starts from the top of the plot and moves to the bottom. From the starting journal, the process evolves towards many other journals depending on outgoing citations listed in the data. The blue out-going links represent the evolution of the diffusion process. The size of the links is proportional to the probability with which a random walker would follow them, i.e., proportional to the number of out-going citations. Note that we plot only the first two steps of the process.

rank	score 1	rank	score 2	change	journal name
1	0.0236	1	0.0465	=	Nature
2	0.0231	3	0.0432	↓	PNAS
3	0.0220	2	0.0434	↑	Science
4	0.0184	34	0.0118	↓	Physical Review Letters
5	0.0165	14	0.0161	↓	Nature Communications
6	0.0147	5	0.0254	↑	The New England Journal of Medicine
7	0.0145	12	0.0187	↓	Plos One
8	0.0144	4	0.0283	↑	Cell
9	0.0142	9	0.0198	=	Nature Medicine
10	0.0141	6	0.0217	↑	Journal of Clinical Investigation

Table 4.2: Ranking of journals according to PageRank. The columns are (from left to right): the ranks and the scores in the *first*-order network, the ranks and the scores in the *second*-order network. The change column contains a red arrow pointing downwards when the journal decreases its position from the rank in the first-order network to the one in the second-order network. Viceversa, when the journal increases its rank position we put a green arrow pointing upwards.

From Fig. 4.4, we immediately see that from both initial journals, we move almost with the same probability to any other journals after two steps. The main difference between Nature Physics (Fig. 4.4(a)) and Plos One (Fig.Fig. 4.4(b)) is in the number of nodes that can be reached after one step: Plos One has more first-order neighbors compared to Nature Physics. This is understandable as Plos One is a multidisciplinary journal, and hence, its papers tend to cite a much wider variety of journals.

In Fig. 4.5, we show the first two steps of a diffusion process evolving on the *optimal*-order network, again starting from Nature Physics (a) and Plos One (b). This time the process starting from the two journals shows more heterogeneity compared to Fig. 4.4. The probabilities of reaching different journals after two steps are visibly different and depend on the initial journal. Such property was utterly lost when aggregating the information on the first-order network. In the next section, we show how this affects the ranking of journals according to PageRank.

rank	score 2	rank	score 1	change	journal name
1	0.0465	1	0.0236	=	Nature
2	0.0434	3	0.0220	↓	Science
3	0.0432	2	0.0231	↑	PNAS
4	0.0283	8	0.0144	↓	Cell
5	0.0254	6	0.0147	↓	The New England Journal of Medicine
6	0.0217	10	0.0141	↓	Journal of Clinical Investigation
7	0.0206	12	0.0138	↓	Nature Genetics
8	0.0198	13	0.0137	↓	Nucleic Acids Research
9	0.0198	9	0.0142	=	Nature Medicine
10	0.0189	14	0.0133	↓	Journal of the American Chemical Society

Table 4.3: Ranking of journals according to PageRank. The columns are (from left to right): the ranks and the scores in the *second*-order network, the ranks and the scores in the *first*-order network. The change column contains a red arrow pointing downwards when the journal decreases its position from the rank in the second-order network to the one in the first-order network. Viceversa, when the journal increases its rank position we put a green arrow pointing upwards.

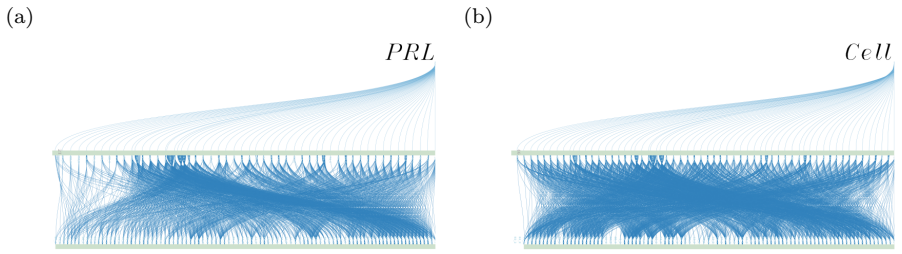


Figure 4.6: Diffusion process computed using the citation network from any starting journal ending in Physical Review Letters (a) and Cell (b).

4.3.2 Ranking journals using the empirical knowledge flow

In Table 4.2, we report the rankings of the top-10 journals among the 94 analyzed ones according to PageRank computed on the first-order network. Additionally, we also report the rank position of these top-10 journals according to PageRank computed on the optimal-order network. From this table, we find that a journal either gains or loses rank positions when computing

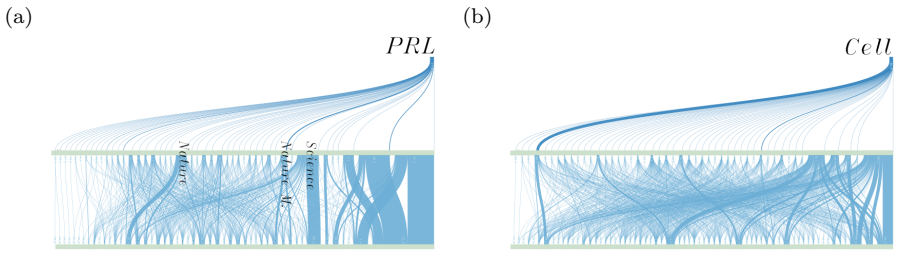


Figure 4.7: Empirical diffusion process computed from any starting journal ending in Physical Review Letters (a) and Cell (b).

its centrality on the optimal-order or it does not move. For example, we find that journals like Nature and Nature Medicine do not change their ranking positions. Whereas, journals like Physical Review Letters (PRL), Nature Communications, and Plos One lose many positions. The most extreme example is PRL that is at the 4th position in the ranking computed using the first-order network and it goes all the way down to 34th position in the ranking obtained with the optimal-order network. For other journals, like Cell, we see an improvement in their raking position.

To understand what drives changes in ranking positions, we use alluvial diagrams. Differently from before, now we do not look at paths starting from a focal node (journal), but we rather consider those paths arriving there. In Fig. 4.7(a), we report the empirical paths of length two that start from any journal and finish in PRL. From this figure, we find that PRL is not as central as one would have guessed by looking at the first-order network Fig. 4.6(a). Precisely, we notice that fewer journals belong to paths reaching PRL, meaning that fewer journals build their knowledge on the one contained in PRL. In addition, most of the paths empirically reaching PRL come from specific journals not particularly central, i.e., few paths go trough them (Fig. 4.7(a)). The only three journals that are central and that are directly connect to PRL are Science, Nature, and Nature Materials. However, from the first two journals, few paths continue to PRL, i.e., thin lines connect Nature and Science to RPL. Only from Nature Materials, we see a more relevant fraction of paths continuing to PRL, i.e., they are connected by a wider and darker line.

In Fig. 4.7(b), we report the empirical paths of length two that start from any journal and finish in Cell. We find that Cell is almost as central as one would have guessed by looking at the first-order network Fig. 4.6(b). Precisely, we notice that Cell is reached from almost any journals in the same way. Moreover, we see that the paths reaching Cell are more and more diverse compared to PRL. Recalling that PageRank scores are relative values and sum up to one, we find that journals like Cell acquire the scores lost by journals like PRL for not being well embedded in the network.

With Fig. 4.7, we have provided a clear visualization of citations paths that represent the empirical flow of knowledge between journals. Additionally, from this visualization, we have provided an intuition of why PageRank scores calculated on the optimal-order network better capture the influence of journals in the empirical flow of knowledge. In the next section, we will continue to show how the optimal-order network allows us to capture better journal properties. In particular, we will investigate topic similarities between journals.

4.4 Classifying the journals

We focus on two problems: 1) how to capture the similarity of two journals and 2) how to assign similar journals to the same category, such as Mathematics or Physics. To develop our similarity measures, we use the clustering algorithm `Infomap` [188]. Recall that this algorithm clusters together nodes that are frequently visited one after the other by a random walker. The random walker visits differently journals following the citation links in the opposite directions. Hence, the trajectory of a random walker reproduces the knowledge flow between journals, and `Infomap` clusters journals where the flow is more stagnant. Note that `Infomap` has already been successfully applied to study the memory effects in citation networks at the journal level [190]. However, compared to this previous work, we apply the clustering algorithm on the optimal order network detected using the procedure presented in [201]. As we have found two as optimal order, we can directly compare the results of [190] with ours.

Order	nodes	links	DOF	p-value
0^{th}	95	94	5109717	≈ 1
1^{st}	94	3038	4385706	≈ 1
$2^{\mathbf{nd}}$	**2689**	**20389**	**3661695**	$\approx \mathbf{0}$
3^{rd}	16029	43651	2999567	≈ 1

Table 4.4: Order detection with `Pathpy` and `Infomap`. For the former, we report the p-values for the Likelihood ratio between to models of increasing order, see eq. 7 in [201] for details). While for the latter we report the Minimum Description Length (MDL), see eq. 1 in [188].

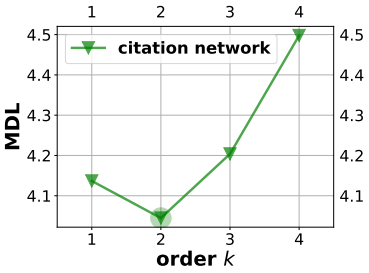


Figure 4.8: Correspondence between MDL and statistical test of `Pathpy`

4.4.1 Clustering journals

We apply `Infomap` on the first, second, third, and fourth-order network. As reported in Table 4.8, we find that `Infomap` finds more tight clusters on the second-order network. This means that both `Infomap` and `Pathpy` suggest the second-order as the most optimal one for both clustering and reproducing the knowledge flows. It is worth noticing that this is an exciting and not trivial congruence. Indeed as shown in [111], `Infomap` and `Pathpy` capture different properties of the data.

We interpret the congruence as follows. In the empirical knowledge flow, the second-order memory effects are not only dominating from a statistical point of view, but they also better capture where knowledge is more often re-combined. We argue that this derives from the presence of multidisciplinary journals that act as intermediaries between specialized journals. Paths of length two in the citation network represent knowledge indirectly exchanged among journals specialized on the same topics via multidisciplinary journals. This phenomenon is so strong that it creates exceptionally tight clusters on the second-order network.

Note that our path abstraction allows us to identify the influence of multidisciplinary journals by capturing the empirical knowledge flow. When considering knowledge flows using paths of length two, we retain information about the origin and final destination of knowledge. Hence, with the path abstraction,

we do not lose information about the journal containing the original knowledge even when this knowledge is crossing multidisciplinary journals. Such information is instead lost when we use a first-order network, i.e., a standard network perspective.

Recognizing the presence of multidisciplinary and specialized journals well explains the congruence between by `Pathpy` and `Infomap`. At the same time, this might not be the only reason why we observe such a congruence. It would be fascinating to understand which are the necessary and sufficient conditions to observe this congruence. However, this goes well beyond the scope of this thesis and left for future research.

In the next section, we build on the clustering obtained by `Infomap` to develop a new similarity measure for journals. We will not discuss the details of the obtained clustering. The analysis of the `Infomap` clustering at journal level has already been done by various works [189, 190]. We will rather use the obtained clustering as input for defining new similarity measures that outperform established ones. Indeed, we will show that our similarity measures are better able to define journal categories compared to established ones.

4.4.2 Similarity of journals

To define a new similarity measure for journals, we use the following idea: two journals are similar when they *often* publish papers containing *similar knowledge*. To define "similar knowledge", we use the fact that new papers cite older ones, and re-build on the knowledge contained in these older papers. Given this, we assume that knowledge contained in citing and cited papers is similar or at least topically similar. By analyzing citations with a path abstraction, we reconstruct paths that proxy the empirical flow of knowledge at the paper level and along these paths, nearby papers contain similar knowledge. Hence, when projecting these paths at the journal level, nearby journals are those publishing papers with similar knowledge.

To define when journals *often* publish papers with similar knowledge, we analyze journals on the *optimal*-order network. On this network, links preserve sub-paths that are statistically more significant, i.e., that appear more fre-

quently in the data. In other words, these sub-paths capture where similar knowledge is often published. Hence, when two or more journals frequently appear close to each other on these sub-paths, we can consider these journals similar.

We capture the similarity of journals by clustering journal pairs with `Infomap` on the *optimal*-order network. In the previous section, we have obtained two as optimal order, and hence, the optimal-order network is a second-order network. On this network, nodes are journal-pairs, and each journal can appear in different nodes. This implies that when clustering nodes on the second-order network, each cluster is a set of journal pairs and each journal can appear in different clusters. With `Infomap` we cluster journal pairs and then, we create for each journal a vector containing the clusters in which it has appeared. Each dimension of this vector represents a cluster, and the vector values represent the number of times a journal was assigned to the different clusters. By this, for each journal, we obtain a feature vector that we can use to measure its similarity to other journals.

To compute journal similarity using their feature vectors, we focus on two similarity measures: Jaccard similarity [100] and weighted Jaccard similarity [191, 206]. These two measures are among the most simple similarity measures that can be computed between sets and are commonly used in many disciplines, ranging from ecology to computer science. Given two sets $M(A)$ and $M(B)$, the Jaccard similarity is the ratio between the size of the intersection and the union of the sets:

$$J(A,B) = \frac{|M(A) \cap M(B)|}{|M(A) \cup M(B)|}\,. \tag{4.1}$$

In our analysis, $M(A)$ is the set of clusters to which journal A belongs. Hence, $J(A,B)$ is the ratio of unique clusters shared between two journals A and B.

Each journal might appear more than once in the same cluster, and hence, we can also use the weighted Jaccard similarity to compare feature vector of journals:

$$Jw(A,B) = \frac{\sum_c \min\left(X(A)_c, X(B)_c\right)}{\sum_c \max\left(X(A)_c, X(B)_c\right)}\,, \tag{4.2}$$

where $X(A)_c$ is the number of times journal A appears in cluster c and the summation is taken over all the detected clusters. Then, Jw gives the fraction of the joint appearances of two journals in the same clusters. The advantage of Jw over J is that Jw uses journal frequencies of appearance in each cluster, while J does not. This allows a more refined comparison when journals are assigned to many different clusters with different frequencies.

We define as our new similarity measures between two journals A and B the $J(A, B)$ and $Jw(A, B)$ where the feature vectors for the journals are obtained from the clustering created by `Infomap`.

Validation procedure. We verify if our similarity measures can identify as similar journals those journals publishing inside the same scientific field. We name scientific fields, such as Mathematics and Physics, *journal categories*. We assume the categories assigned by Clarivate Analytics (CA) to journals as the correct categories of journals [7]. In other words, we assume the CA journal categorization as the ground truth for our data in our validation procedure. Our validation procedure is divided in three steps.

First, we match journal names from CA with names coming from Google Metrics. We were able to match 85 journal names out of the 94 under analysis. The names not matched are nine arXiv journals as these are not present in the CA database. We report in Table 4.5 the names of the CA categories together with their sizes. We see that among 22 categories only 12 contain a paper belonging to the top-85 matched journals.

Second, we calculate two quantities: the *in*- and *out*-category similarity. The *in*-category similarity is the sum of similarities between pairs of journals belonging to the *same* category. We can express this quantity with the following formula:

$$\tilde{S}^m_{in}(\mathcal{C}) = \sum_{c_\alpha \in \mathcal{C}} \sum_{A,B \in c_\alpha, A \neq B} S^m(A, B) \tag{4.3}$$

[7]Note that Thomson Reuters previously owned CA. Their journal categorization can be retrieved at `http://ipscience-help.thomsonreuters.com/incitesLiveESI/10678-TRS.html`

Category	Clarivate	Top-100 from Google
AGRICULTURAL SCIENCES	344	0
BIOLOGY & BIOCHEMISTRY	431	4
CHEMISTRY	531	10
CLINICAL MEDICINE	1928	22
COMPUTER SCIENCE	394	0
ECONOMICS & BUSINESS	583	2
ENGINEERING	854	4
ENVIRONMENT/ECOLOGY	355	2
GEOSCIENCES	414	0
IMMUNOLOGY	165	4
MATERIALS SCIENCE	363	7
MATHEMATICS	485	0
MICROBIOLOGY	124	0
MOLECULAR BIOLOGY & GENETICS	302	9
MULTIDISCIPLINARY	52	6
NEUROSCIENCE & BEHAVIOR	330	4
PHARMACOLOGY & TOXICOLOGY	273	0
PHYSICS	313	9
PLANT & ANIMAL SCIENCE	795	0
PSYCHIATRY/PSYCHOLOGY	633	0
SOCIAL SCIENCES, GENERAL	1977	0
SPACE SCIENCE	54	2
TOTAL	11700	85

Table 4.5: Categories of Clarivate Analytics with their size in their data and our data.

where $\tilde{S}^m_{in}(\mathcal{C})$ is the *in*-category similarity according to similarity measure m, $\mathcal{C}$ is a journal categorization, i.e. $\mathcal{C} = \{c_1, c_2, ..\}$, $c_\alpha = \{A, B, ..\}$ is a set containing the various journal assigned to category-α and $S^m(A, B)$ is the similarity score between A and B according to the indicator m. The *out*-category similarity is the sum of similarity between pairs of journals belonging to *different* categories. We can write this as

$$\tilde{S}^m_{out}(\mathcal{C}) = \frac{1}{2} \sum_{A \in c_\alpha} \sum_{B \in c_\beta, \alpha \neq \beta} S^m(A, B)\,, \tag{4.4}$$

$\tilde{S}^m_{in}(\mathcal{C})$ is the *out*-category similarity according to the similarity measure m. The $\tilde{S}^m_{in}(\mathcal{C})$ and $\tilde{S}^m_{out}(\mathcal{C})$ respectively quantify how similar are journals belonging to the same or to different categories. In other words, given a journal categorization $\mathcal{C}$, a good similarity measure m should have large $\tilde{S}^m_{in}(\mathcal{C})$ and small $\tilde{S}^m_{out}(\mathcal{C})$.

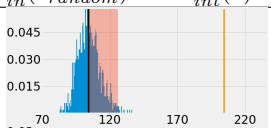
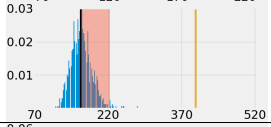
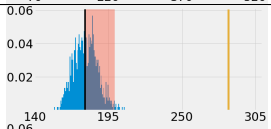
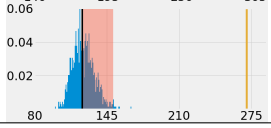
Metrics		$\tilde{S}^m_{in}(\mathcal{C}_{random})$ and $\tilde{S}^m_{int}(\mathcal{C})$
1st Order Network	Bibliographic (Out-sim)	
	Cosine Co-occ. (In-sim)	
Optimal Order Network	Jaccard	
	Jaccard (weighted)	

Table 4.6: We compare the aggregated in-cluster similarity score coming from the different similarity measures. For each similarity measure, we plot their *observed* in-cluster score (in yellow) and the *distribution* of in-cluster scores simulated by randomizing the categorizations of the journals (in blue). For these distributions, we also report their average values (in black) and their (right side of the) 95% confidence intervals (red shaded area). We find that for all similarity measures, the observed in-cluster score is much bigger than the simulated scores and hence, all measures are good in detecting similarities among journals.

Third, we compute the $\tilde{S}^m_{in}(\mathcal{C})$ and the $\tilde{S}^m_{out}(\mathcal{C})$ for our new two indicators and for two established ones. The two established indicators that we consider are bibliographic coupling B_c and cosine similarity based on co-citations C_s. For the definition of these two indicators see Sect. 2.3.2. In Table 4.6, we report the $\tilde{S}^m_{in}(\mathcal{C})$ computed using the journal categorization $\mathcal{C}$ coming from CA for the four similarity measures, $m = B_c$, C_s, J and Jw. To understand if the computed values of $\tilde{S}^m_{in}(\mathcal{C})$ are significantly large, we show how distant they are from the distribution of $\tilde{S}^m_{in}$ coming from a sample of randomized categorizations $\{\mathcal{C}_{random}\}$. These are obtained by assigning the journals to random categories while keeping the size of the categories equal to the true ones. For the four analyzed measures, we find that their $\tilde{S}^m_{in}(\mathcal{C})$ are far away

(bigger) from the distribution of $\tilde{S}^m_{in}(\mathcal{C}_{random})$. This finding means that the high values of $\tilde{S}^m_{in}(\mathcal{C})$ could not be obtained with a random categorization. Hence, all the similarity measures can capture the similarity between journals belonging to the same category. We have performed the same analysis also for $\tilde{S}^m_{out}(\mathcal{C})$ and we have obtained similar results. For the four analyzed measures, the $\tilde{S}^m_{out}(\mathcal{C})$ are statically significant smaller compared to the values coming from the distribution of $\tilde{S}^m_{out}(\mathcal{C}_{random})$ (not shown).

With the above procedure, we have verified that our indicators capture similarity values that are consistent with the journal categorization of CA. In addition, we have shown that their capacity to capture similarities between journals is *qualitatively* comparable to other established measures. How can we *quantitatively* compare our indicators with the established ones? Which is the best indicator capturing journal similarity?

J and Jw outperform established indicators. To establish which of the analyzed indicators m better captures journal similarity, we compare them using their $\tilde{S}^m_{in/out}(\mathcal{C})$. Note that the absolute values of $\tilde{S}^m_{in/out}(\mathcal{C})$ cannot be directly compared as they come from different measures that have different properties. Indeed, the distributions of $\tilde{S}^m_{in/out}(\mathcal{C}_{random})$ are centered at different values and their 95% confidence bounds have different sizes (see Tab. 4.6). Therefore we use a *train-test prediction* approach to compare the different measures. In other words, we verify if we can predict the category of journal A (test) using its similarities with other categorized journals (train). To do this, we divide the data in train-test and use a simple greedy algorithm to predict the categories of the journals in the test group. For each similarity measure, the greedy algorithm performs the following steps:

1. it chooses a journal A from the test group
2. for all possible categories c_α, it computes the ratio $f_\alpha = \tilde{S}^m_{in}/\tilde{S}^m_{out}$ by considering $A \in c_\alpha$
3. it assigns journal A to the category-α^* with highest ratio, i.e. $\alpha^* = argmax_\alpha f_\alpha$

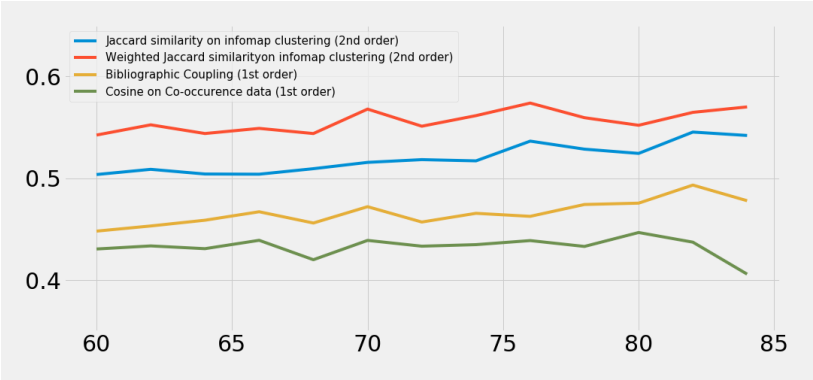


Figure 4.9: Precision of the different similarity measures in function of the size of the train group. For each training size, we have divided the journals randomly in train and test 500 times. The reported precision is the average one.

4. it removes the classified journals from the test group and repeats all steps until all journals are classified

In Fig. 4.9, we report the precision of the measures in predicting categories. For each measure, we compute its precision, i.e., the ratio between the number of correctly predicted categories over the size of the test group. The precision of the greedy algorithm is higher when using the weighted Jaccard on the second order network. This means that among the analyzed indicators, Jw better captures the similarities between journals. Additionally, we find that both the indicators computed on the second-order network outperform the ones computed on the first-order network.

What do we learn? From the comparison of the different indicators, we have found that we better capture journal similarity when considering the optimal-order network. Now we look for which particular pairs of journals we have an information gain. In other words, what new or different information we obtain by using Jw? To answer this question, we compare the similarity scores of journal pairs coming from different indicators. Since the absolute scores should not be directly compared, we compare their percentiles. Given

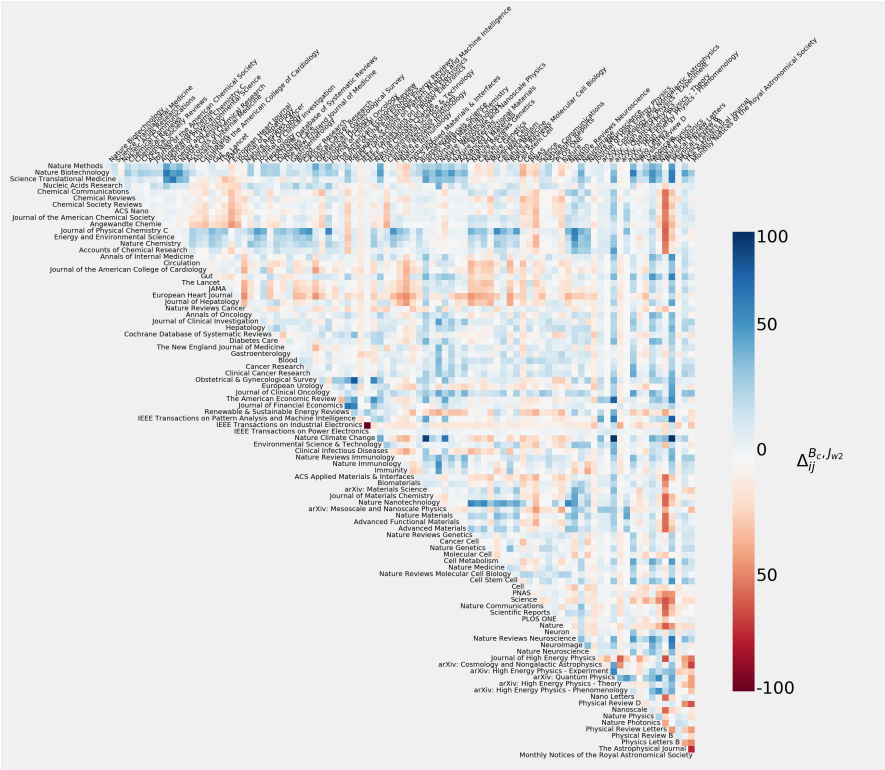


Figure 4.10: $\Delta^{(m_1,m_2)}$ between all journal pairs where m_1 is the bibliographic coupling B_c and m_2 is Jaccard similarity computed using the infomap clustering on the optimal order network

a measure m, we have a score for each pair of journals $S^m(A,B)$ and hence, a distribution of similarity scores S^m. For each similarity score $S^m(A,B)$, we compute its percentile $p^m(A,B)$ in S^m. Note that we have decided to use the percentile instead to compute other values, such as the z-score, as the distribution of similarity scores are bounded between [0,1] and skewed (not shown). After calculating the percentiles for two different measures m_1 and m_2, we take their difference:

$$\Delta^{(m_1,m_2)}(A,B) = p^{m_1}(A,B) - p^{m_2}(A,B) \tag{4.5}$$

overestimated		
journal-A	journal-B	$\Delta^{(B_c,J_w)}$
arXiv: Quantum Physics	Nature Climate Change	87.0
Nature Climate Change	Physical Review B	86.0
arXiv: Materials Science	Nature Climate Change	81.0

Table 4.7: Pairs of journals for which Bibliographic coupling has mostly *over*-estimated the similarity.

overestimated				
journal-A	journal-B	$S^{B_c^{mult}}$	$S^{B_c^{same}}$	$S^{B_c^{others}}$
arXiv: Quantum Physics	Nature Climate Change	0.3838	0	0
Nature Climate Change	Physical Review B	0.3611	0	0
arXiv: Materials Science	Nature Climate Change	0.2841	0	0

Table 4.8: Breakdown of the Bibliographic coupling for the mostly *over*-estimated pairs of journal by using Eq. (4.6).

In Fig. 4.10, we show the $\Delta^{(m_1,m_2)}(A,B)$ between the percentile scores of bibliographic coupling B_c and weighted Jaccard similarity Jw as a heat map. Each cell represents a journal pair and cell color the percentile difference: When $\Delta^{(B_c,J_w)}(A,B) \gg 0$, we use an intense blue color and it means that the bibliographic coupling B_c was *over*-estimating the similarity between the journals A and B. While when $\Delta^{(B_c,J_w)}(A,B) \ll 0$, we use an intense red color and it means that B_c was *under*-estimating the similarity between the journals A and B.

To understand why there are big differences in the similarity scores computed by the different measures, we focus on the most *over*-estimated pairs (see Table 4.7). For each pair (A,B) we split its bibliographic coupling score in three terms: a first term coming from citations belonging to journals in the same categories of A and B, B_c^{same}, a second term from citations belonging to multidisciplinary journals, B_c^{mult}, and a third term with the rest, B_c^{others}. Hence, for every journal pair (A,B),

$$S^{B_c}(A,B) = S^{B_c^{same}}(A,B) + S^{B_c^{mult}}(A,B) + S^{B_c^{others}}(A,B)\ . \tag{4.6}$$

We find that the bibliographic coupling scores for the most *over*-estimated pairs are dominated by the term $S^{B_c^{mult}}(A,B)$ (see Table 4.8). This means that dissimilar journals, like "arXiv: Quantum Physics" and "Nature Climate

Change" were considered similar only because multidisciplinary journals were citing both of them. Note that by using a path abstraction to reproduce the knowledge flow and measure journals similarity, we do not commit such wrong judgments.

4.4.3 Multidisciplinary index

Given the statistical evidence that the second-order network optimally captures knowledge flows (see Sect. 4.3.1), we now propose a multidisciplinary index for journals using this evidence and an idea proposed by [190]. In this work, the authors suggest that we can detect multidisciplinary journals by using the `Infomap` clustering. Here, we do the following: first, for each journal A we count the number of clusters to which A is assigned by `Infomap` in the second-order network, and then, we rank journals according to this number. In other words, the multidisciplinarity index of journal A is $|M(A)|$, i.e. the size of the feature set $M(A)$ defined in Sect. 4.4.2. Recall that number of clusters to which a given journal is assigned is the number of "stagnant flows" in which this journal appears. This means that we quantify the number of distinct knowledge flows to which a given journal contributes, and hence, we are quantifying how diverse is the knowledge that its papers foster.

In Table 4.9, we report the 20 most multidisciplinary journals according two this procedure. We find that 6 out of the top-10 are the journals belonging to the MULTIDISCIPLINARY category. This is a first good result as it confirms that the via the method suggested by [190] we can detect multidisciplinary journals.

Ranking bias. To verify that the ranking obtained in Table 4.9 is not biased in favor of any category, we use the approach developed in Chap. 3. To do this, we test whether the ranking vector containing the top-20 journals favor or disfavor any category. Note that in these top-20 journals, we expect to find the six multidisciplinary journals in our sample (that we actually find) and other 14 journals randomly sampled from the remaining. Thus, we sample 14 journals without replacement various times and compute the Mahalanobis distance ($d_{\mathcal{M}}$) between the sampled ranking vectors and the expected one. In

rank	score	modules	journal name (category)
1	1.0	28	Plos One (MULTIDISCIPLINARY)
2	2.0	26	Scientific Reports (MULTIDISCIPLINARY)
3	3.0	23	Nature Communications (MULTIDISCIPLINARY)
4	4.0	20	PNAS (MULTIDISCIPLINARY)
5	5.5	19	Clinical Cancer Research (CLINICAL MEDICINE)
6	5.5	19	Nature Medicine (MOLECULAR BIOLOGY & GENETICS)
7	7.5	18	Biomaterials (MATERIALS SCIENCE)
8	7.5	18	Science (MULTIDISCIPLINARY)
9	9.5	17	Nature (MULTIDISCIPLINARY)
10	9.5	17	Nucleic Acids Research (BIOLOGY & BIOCHEMISTRY)
11	11.5	16	Cancer Research (CLINICAL MEDICINE)
12	11.5	16	Nature Methods (BIOLOGY & BIOCHEMISTRY)
13	14.0	15	Blood (CLINICAL MEDICINE)
14	14.0	15	Gastroenterology (CLINICAL MEDICINE)
15	14.0	15	The New England Journal of Med. (CLINICAL MEDICINE)
16	19.0	14	ACS Nano (CHEMISTRY)
17	19.0	14	Cell (MOLECULAR BIOLOGY & GENETICS)
18	19.0	14	Cell Stem Cell (MOLECULAR BIOLOGY & GENETICS)
19	19.0	14	Chemical Reviews (CHEMISTRY)
20	19.0	14	Hepatology (CLINICAL MEDICINE)

Table 4.9: Multidisciplinary Ranking of scientific journals. Ties are resolved using alphabetic order. The ranking score is obtained by the average position of the journals with the same number of modules. For example we have two journals with 19 modules in rank position 5 and 6, then we assign to both journals score 5.5..

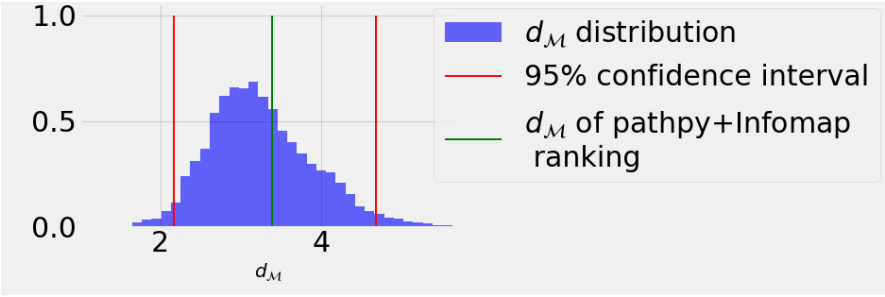


Figure 4.11: In blue, the distribution of the $d_{\mathcal{M}}$ calculated between the ranking vectors coming from the unbiased selection process and the expected vector $\vec{\mu}$ computed using the methodology described in Chap. 2. In red, we provide the 95% confidence interval of the distribution. In green, we report the $d_{\mathcal{M}}$ between the ranking vector coming from our multidisciplinary score and $\vec{\mu}$.

rank	score	size	name
1	4.5	6	MULTIDISCIPLINARY
2	16.2	4	BIOLOGY & BIOCHEMISTRY
3–4	21.3	22	CLINICAL MEDICINE
3–4	21.3	9	MOLECULAR BIOLOGY & GENETICS
5	22.9	7	MATERIALS SCIENCE
6	23.9	4	NEUROSCIENCE & BEHAVIOR
7–9	24.0	10	CHEMISTRY
7–9	24.0	4	ENGINEERING
7–9	24.0	2	ECONOMICS & BUSINESS
10	24.5	2	ENVIRONMENT/ECOLOGY
11	24.6	4	IMMUNOLOGY
12	24.8	9	PHYSICS
13	25.8	2	SPACE SCIENCE

Table 4.10: Multidisciplinary Ranking of scientific categories according to the Multidisciplinary score of its journals. We use Borda counting to aggregate the scores of the different journals.

Fig. 4.11, we report the distribution of $d_{\mathcal{M}}$ of the sampled vectors and of the ranking obtained with our multidisciplinarity index. We find that the $d_{\mathcal{M}}$ of our ranking follows inside the 95% band of the distribution of the unbiased rankings. This implies that the ranking generated by our indicators is not biased, and hence, it does not favor or disfavor any categories.

We further use our multidisciplinarity index to evaluate which journal category is more multidisciplinary. To do this, we aggregate ranking positions of journals per category using a Borda count approach. We assign to each category a score equal to the average ranking positions of journals belonging to the same category. Then, rank journal categories in increasing order such that the category with a lower (average) score is at the top of the ranking. By this, the most multidisciplinary category is the one that has more journals in the top positions in the multidisciplinarity ranking obtained in the previous paragraph.

We report our multidisciplinary (Borda) ranking for categories in Table 4.10. We find that at the top of the ranking the category MULTIDISCIPLINARY. This result is expected. We also find that all the other categories have higher scores compared to MULTIDISCIPLINARY, but their scores are close to each other. This is a good indication that the multidisciplinary index can separate multidisciplinary and specialized categories. The only exception is BIOLOGY

& BIOCHEMISTRY at the second position in the ranking that has a score close to 16, while the remaining categories have scores bigger than 21. This is something slightly unexpected. One could have argued that scientific categories doing more fundamental research, like CHEMISTRY and PHYSICS, should have appeared at the top of the ranking. Instead, we find that these two disciplines are the bottom of the ranking with scores close to 24.

4.5 Conclusion

Increasing attention has been given to data in order to govern science [94]. This has produced the need for developing new and more sophisticated measures for quantifying scientific performance. In particular, measures have been obtained by combining citation analysis and network theory. In this chapter, we have shown how a naive combination of the two disciplines produces misleading and, in some cases, even wrong results. By assuming that papers are knowledge artifacts and that citation links represent flows of knowledge, the citation network at the paper level can be used to study the flow of knowledge among these artifacts. However, when projecting citations from the paper to the journal level, we lose information about the empirical knowledge flows.

To overcome the projection problem, we have provided a solution based on the path abstraction [190, 201]. In addition, to recover methods and tools from network analysis, we have used the statistical test developed in [201] allowing us to select an optimal higher-order network to represent our citation data. We have found that a second-order network optimally represents the 150 000 citation links among 94 top-journals listed in our sub-sample of the `MAG17` dataset. Note that we have recently used a more recent version of `Pathpy` allowing us to analyze a bigger sub-sample of citations. We analyzed more than 240 000 citations among our 94 journals and have detected that three as optimal-order. In other words, by increasing the number of citations, we have more data that justify even more a path abstraction.

By obtaining a second-order network as an optimal-order network in our sub-sample, we learn that the analyzed data contains many identical paths and sub-paths of length 2. These paths are triples of papers belonging to journals

always citing each other in a specific order. These frequent paths break the transitivity assumption necessary for applying standard network methods to analyze the empirical knowledge flow on the journal citation network. In other words, any past applications of algorithms or metrics using powers of the adjacency matrix of the journal citation network might have provided wrong results.

By relying on a naive projection of citation links from the paper to the journal level, one would overestimate how much knowledge is exchanged among journals. To do this, we have used alluvial diagrams to visualize the diffusion processes (of knowledge) evolving on the citation network from and to specific journals. With these diagrams, we have shown that the empirical knowledge flow is captured by a higher-order topology containing fewer paths compared to what one would have obtained from the ordinary network perspective.

Additionally, we have analyzed whether the overestimation of knowledge exchange is stronger for some journals. To capture this, we have computed the PageRank scores of journals on both the first and the optimal-order network. We have observed that some journals, like PRL, lose rank positions when moving from the first-order to the optimal-order network, whereas others, like Cell, move up in the ranking. At the same time, by inspecting the alluvial diagrams for PRL and Cell, we find that the number of paths to which both journals belong is overestimated in the first-order network. However, for Cell, the difference between the alluvial diagrams in the first-order and the optimal-order network is smaller. This means that the moving up of a journal, like Cell, in the ranking computed using the optimal order does not imply an *absolute* higher contribution to the empirical knowledge flow. It rather implies a *relative* higher contribution, i.e., there are other journals, like PRL, for which one overestimates even more their contribution the knowledge flow when using a first-order network.

In Sect. 4.4, we have shown that we can construct more informative similarity measures for journals by using the optimal-order network. In our case, the optimal order is two, meaning that nodes of the optimal-order networks are pairs of journals. By clustering nodes with `Infomap` on the optimal-order network, we have found that journals are assigned to multiple clusters. For

every journal, we have grouped their assigned clusters in different sets. Using these sets, we have computed the weighted Jaccard similarity between every pair of journals and compared this to other established citation-based similarity measures: Bibliographic coupling and cosine distance. We have performed the comparison by checking which measure was better able to identify journal categories (MULTIDISCIPLINARY, PHYSICS, etc.) assigned by CA. We find that our combination of `Infomap` on the optimal-order network with the weighted Jaccard similarity outperforms the other analyzed similarity measures in recovering journal categories.

Finally, we have created a multidisciplinary indicator for journals and journal categories. Starting from an idea of [190], we have first created a ranking of journals from more to less multidisciplinary by counting the number of clusters to which `Infomap` assigns each journal. Then, by aggregating journal ranking positions with the Borda's method, we have produced a multidisciplinary score for each category and ranked them using this score. We find that most categories are not multidisciplinary, but rather specialized and similar in having most of the empirical knowledge flow running inside them. Indeed, even the second-ranked category, BIOLOGY & BIOCHEMISTRY, has a score that is four times bigger compared to the first ranked category (MULTIDISCIPLINARY) that contains only multidisciplinary journals. This means that on average journals belonging to BIOLOGY & BIOCHEMISTRY obtain positions that are four times lower (less multidisciplinary) compared to journals belonging to the MULTIDISCIPLINARY category.

To conclude, in this chapter, we have shown how to retain a network perspective when studying citation data at the journal level. This was not a trivial task as we had to combine two novel modeling approaches based on path abstraction and higher-order network models. However, we find that this procedure is necessary to preserve the empirical knowledge flow traversing the journals. Furthermore, it permitted to develop new refined methods to determine how journals are influent, similar, and multidisciplinary. With all this, we not only provide a new perspective to use citation analysis at the journal level but also better tools to support research evaluators and administrators in the challenging tasks of assessing scientific performance and governing science.

Chapter 5

Quantifying knowledge exchange in R&D networks: A data-driven model

Summary

We propose a model that reflects two important processes in R&D activities of firms, the formation of R&D alliances and the exchange of knowledge as a result of these collaborations. In a data-driven approach, we analyze two large-scale data sets extracting unique information about 7500 R&D alliances and 5200 patent portfolios of firms. This data is used to calibrate the model parameters for network formation and knowledge exchange. We find that R&D alliances have a duration of around two years and that the subsequent knowledge exchange occurs at a very low rate. Hence, we find that firm's position in the knowledge space is rather a determinant than a consequence of its R&D alliances. From our data-driven approach we also find model configurations that can be both realistic and optimized with respect to a collaboration efficiency measure $\hat{\mathcal{C}}_n$. This is a new measure, that takes also in account the effort of firms to maintain concurrent alliances, and is evaluated via extensive computer simulations. [1]

[1] Based on [233]

5.1 Introduction

The last three decades have been characterized by a growing number of inter-firm alliances, aimed at Research and Development (R&D) purposes. Albeit this phenomenon has especially affected highly technological industries such as IT, Pharmaceuticals or Medical Supplies [4, 85], all industrial sectors have simultaneously experienced an increased number of such alliances [223].

From a theoretical point of view, it has been shown that firms engage in alliances for several reasons. They can gain access to more and diverse assets [46, 121]. Next, alliances foster the exchange of knowledge between firms: by joining their technological resources, firms can actually enlarge their knowledge bases faster than they could do individually [17, 139, 184]. Finally, firms can share the costs and risks of a project, especially when this is expensive or with uncertain outcome [87]. All of these aspects result in a learning process of the involved firms, making R&D alliances an important part of every firm's knowledge management strategy.

The focus of the present study is indeed such a learning process, which we model as a mutual exchange of knowledge occurring after the establishment of an alliance between two firms. In particular, we develop an agent-based model to investigate the determinants leading to the formation of inter-firm R&D collaborations and the subsequent emergence of an R&D network. Additionally, we estimate the *performance* of such networked systems, in terms of explored technological space. Note that to do this, we will discard differences in the collaboration strategies that firms adopt across industrial sectors [199, 227]. At the same time, with this simplification, we can shift our focus from the specific collaboration strategies of firms to the collective effect that these collaborations have on the technological positions of firms.

The approach that we adopt in our study can be defined as *data-driven modeling*. Starting from the empirical evidence, we design a set of realistic and theoretically grounded microscopic interaction rules, which we incorporate in an agent-based model; next, we implement the model through computer simulations, followed by calibration and validation against empirical data. The fine-tuning of the model parameters gives us not only a deep understanding

of the system under examination, but also an indication on how to optimize it. The model that we develop here is based on previous empirical findings [88, 187, 223], and combines two existing agent-based models [224, 226], in order to reproduce both the alliance formation and the knowledge exchange process in an R&D network.

5.1.1 Theoretical foundations: knowledge exchange in inter-firm R&D networks

Our agent-based model follows a number of extant works on bounded confidence and continuous opinion dynamics [10, 49, 50, 80, 92], in particular applied to innovation networks [16, 56]. In the wake of this previous work, and similar to the model proposed by Tomasello *et al.* [226], we assume that the collaborating agents are characterized by an evolving knowledge basis, that is affected by the set of alliances in which are involved. However, differently from the studies that have been done so far, our model does not focus on the formation of consensus clusters – see Axelrod [10], Schweitzer and Behera [208] in the case of social systems, or Fagiolo and Dosi [54] for technology islands, but on the exploration of a *k*nowledge space (defined below). In addition, our study does not consider the network of R&D alliances as fixed, but it assumes a dynamically evolving R&D network, whose topology corresponds to those of empirically observed networks [see 84, 223].

The knowledge-based view of the firm [56] assumes that every company is endowed with a knowledge basis that uniquely identifies its resources and capabilities. In other words, a firm can always be associated with a vector consisting of several components [192], each of which represents its level of knowledge in a given area. These vectors can in turn be associated with a metric *knowledge space* in which the collaborations occur. Thus, every firm occupies a point in this multi-dimensional space, whose coordinates are given by its knowledge vector. Such an approach is similar to a more general model [10], proposed in the broader context of social influence. The concept of a metric knowledge space has already been used in one [80], and two dimensions [16, 54]; here, we generalize this approach to metric spaces of arbitrary dimensionality.

On the other hand, R&D alliances have been conceptualized by several studies [77, 79, 139, 157] as a means to exchange technological knowledge among firms, and such an idea is at the heart of several agent-based models [43, 74, 172, 174]. In these models, agents' knowledge bases become more similar over time, as a consequence of R&D collaborations. The speed at which the agents approach each other in the knowledge space represents one of the aspects of this family of models, and our work is no exception. Besides, we rely on the assumption that knowledge spillovers occurring in a R&D alliance cause the partners to exchange knowledge along every dimension of their knowledge bases, not limiting the transfer to a specific R&D project that they have in common [16]. Note this assumption makes our model different to other ones, like the one presented in [173], where firms decide to have collaborations to develop specific products and hence, acquire only specific knowledge from their partners. In our data-driven approach, introducing firms' preferences to specific knowledge dimensions is not justified as we do not have data about the reason behind the R&D alliances. Thus, we study a scenario in which two partner firms approach with respect to all dimensions of the knowledge space.

Finally, we aim at studying the *performance* of the whole collaboration network as a function of the relevant model parameters. To quantify it, we propose a measure that takes into account the global knowledge exploration of the systems. I.e., it takes into account the distances in knowledge space traveled by all agents during the evolution of our simulated R&D network. In our model, we consider that the knowledge exploration itself is represented by the motion in the knowledge space, which is fully captured by such a measure. The underlying assumption is that a throughout exploration of the knowledge space is beneficial for the R&D network, in that it allows the agents to come in contact with many technological opportunities, potentially leading to more frequent innovations [54]. Precisely, we make use of an existing performance indicator [226] and refine it by taking into account the actual number of active collaborations in the system, in order to obtain a more reliable measure.

5.1.2 Theoretical foundations: formation of inter-firm R&D networks

The extant literature on R&D networks has shown that two crucial types of mechanisms drive the formation of new R&D alliances [186]: *endogenous* mechanisms and *exogenous* mechanisms. The endogenous mechanisms depend on firms' social capitals which describe the firms' positions in the network, while the exogenous mechanisms are affected by firms' technological and commercial capitals. Here, we refer to an alliance as "endogenous" if it involves a partner that belongs already to the R&D network. While if it involves a partner that does not belong to the R&D network, we refer to the alliance as "exogenous".

Typically, empirical and theoretical studies have focused on the mechanisms driving endogenous and exogenous alliances separately, also called "network endogeneity" [65, 83, 168, 245] and "exogenous partner selection" [32, 42, 185]. However, to explain the observed empirical R&D network both types of mechanisms are needed. As matter of fact, network endogeneity by itself would produce more and more centralized network over time, which does not occur in the real R&D network [223]. On the other hand, a purely exogenous partner selection would lead to regular network topologies, which also does not occur (a prominent example is represented by the "monogamous" networks analyzed by Tomasello *et al.* [226]). A notable exception is the agent-based model developed by Tomasello *et al.* [224], which incorporates both endogenous and exogenous rules of alliance formation and successfully reproduce the structure of a real R&D network. In fact, the model permits to tune the weight of both endogenous and exogenous mechanisms for alliance formation, and to test the outcome against real data.

Inspired by these works, the agent-based model that we develop in the present study includes all the microscopic rules introduced in Tomasello *et al.* [224], and combines them with the knowledge exchange rules briefly discussed above. Our model allows us to modulate the weight of both endogenous and exogenous mechanisms for alliance formation, and to study the knowledge exchange in R&D networks.

5.1.3 Our contribution

As mentioned, we combine, and extend, two existing agent-based models in a straightforward, yet effective, manner. The model introduced by Tomasello *et al.* [226] represents a first attempt to investigate the process of knowledge exchange occurring in a dynamic collaboration network; it has identified a mechanism of volatile alliances to help the collaborating agents better explore a knowledge space, using the approximation of monogamous (i.e. sparse) collaboration networks. On the other hand, the model developed by Tomasello *et al.* [224] can realistically reproduce the complex topology of real R&D networks, but without considering the effect of alliances on the firms' knowledge positions.

The agent-based model we introduce here constitutes an important step toward a general framework that combines two dynamic processes, the formation of alliances and the knowledge exchange in collaboration networks. The microscopic interaction rules of our model and its calibration involve a two-step procedure that can be described as follows. The firms form R&D collaborations based on their network features and their social capital; the model parameters related to these mechanisms are estimated through a comprehensive inter-firm alliance data set. Next, we assume that the formation of each network link causes a process of knowledge exchange between the involved firms, which consequently approach in the knowledge space; the model parameters related to this mechanism are estimated through a second data set on firm patents. Remarkably, the underlying knowledge space that we consider in our study is defined by real patent classes, allowing for a precise quantification of every firm's technological position. In this chapter, we also investigate how the dimensionality of the knowledge space impacts our results.

Our findings point out a predominance of the endogenous network mechanisms (over the exogenous ones) for the alliance formation; in other words, previous network structures and alliance history matter when selecting new collaboration partners. Next, we find that real R&D alliances have a duration of around two years, and that the subsequent knowledge exchange between the partners occurs at a very low rate. Most of the alliances, indeed, have no consequence on the partners' knowledge position: this suggests that a firm's position – eval-

uated through its patents – is rather a determinant than a consequence of its R&D alliances. Finally, we investigate the performance of such a network in terms of explored knowledge trajectories, and we check whether the real R&D network under examination maximizes our proposed performance indicator. Interestingly, we find that this is the case: effective policies to obtain an optimized collaboration network – as suggested by our model – would incentivize shorter R&D alliances and higher knowledge exchange rates.

The rest of the chapter is organized as follows. Section 5.2 presents the data sets and the methodology used to build the network, as well as to measure the firms' knowledge positions. Section 5.3 describes all the microscopic interaction rules defining our agent-based model. Sections 5.4.1 and 5.4.2 present the results of our computer simulations and the model calibration on the alliance and the patent data sets, respectively. In Section 5.5, we introduce a quantification of the collaboration efficiency and study the optimality of the real R&D network under examination. Finally, Section 5.6 concludes.

5.2 Data and Methodology

5.2.1 Network reconstruction, activities and patents

We define an R&D network as a set of nodes, or agents (the firms), and links (the alliances between them). By R&D alliance (or collaboration), we refer to an event of partnership between two firms that can span from formal joint ventures to more informal research agreements, specifically aimed at research and development purposes. To detect such events, we use the SDC Platinum database, provided by Thomson-Reuters [221], that reports all publicly announced alliances, from 1984 to 2009 between several kinds of economic actors (including manufacturing firms, investors, banks and universities). In our network representation, we draw an undirected link connecting two nodes every time an alliance between the corresponding firms is announced in the data set. When an alliance involves more than two firms (consortium), all the involved firms are connected pairwise, resulting into a fully connected clique. This procedure is consistent with a previous empirical study [223], where there is

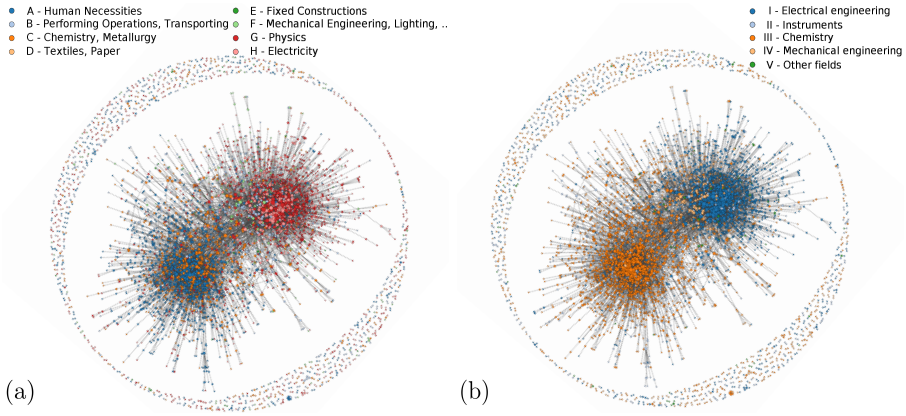


Figure 5.1: The R&D network: each node is a firm and its color refers to the domain where the firm has filed more patents between 1984 and 2009. For figure (a) we used the main 8 IPC-sections to classify the patents, while for (b) we used the main 5 areas from ISI-OST-INPI classification scheme. For a discussion about the colors of the nodes see Sect.5.2.2. We use the layout algorithm of [64] for both networks.

no conceptual difference between a consortium and a "standard" two-partner alliance, which is only a special case of it (and can be thought of as a fully connected clique of size 2). Fig. 5.1 shows a visualization of the time aggregated R&D network, where each node is a firm and links are alliances listed in the above mentioned data set.

A quantity that we measure directly from the data prior to the implementation of our agent-based model is the firms' *activity* distribution.[2] The activity expresses the probability that a firm takes part in any alliance event occurring in a given time window. For the calibration of the present model we use the overall firm activity, measured on the entire observation period of the data set. We define such activity a_i of firm i as the number of alliance events e_i involving firm i divided by the total number of alliance events E involving *any* firm reported in the data set. We then assign such empirical activities a_i to the agents in our computer simulations.

[2]For a more detailed definition and more empirical examples on agents' activity in collaboration networks see Tomasello *et al.* [224] and its Supplementary Information.

The SDC Platinum database [221] reports approximately 672'000 publicly announced alliances in all countries with a granularity of 1 day. We apply two filters: first, to select only the alliances characterized by the "R&D" flag; with this, we obtain a list of 14'829 alliances, connecting 14'561 firms. Second, we keep in our network representation only firms that have a corresponding entry in the patent data set such that we can determine their knowledge positions. The patent database used is the Patent Citations Data by the U.S.A. National Bureau of Economic Research (NBER), that contains detailed information on patents granted in the U.S.A. and other contracting countries, from 1971 to present. Obviously, we select only the entries that have a match with the SDC alliance data set, both with respect to assignees and time period, thus obtaining a total of around 1'400'000 listed patents. Every patent is associated with one or more assignees and with an International Patent Classification (IPC) class. Companies are associated with a unique identifier, and a relatively big part of them (5'168 firms, precisely) are matched to the SDC alliance data set. These firms take part in 7'417 distinct R&D alliances.

5.2.2 Firms positions in knowledge space

Classification schemes In this chapter we we use – and compare – different approaches to determine the knowledge position of a firm. Both approaches compute the shares of patents of a firm with respect to two different classification schemes, the Industrial Patent Classification (IPC) and the Fraunhofer ISI, Observatoire des Sciences et des Techniques (OST) and French patent office (INPI) classification (ISI-OST-INPI). These classifications differ in the number of classes taken into account, which will correspond to the dimensionality of the knowledge space in which the firms are located. IPC operates on 8 dimensions, while ISI-OST-INPI considers 35 dimensions. More details are given in the following.

The IPC, introduced in 1971 by the *Strasbourg Agreement*, is a hierarchical system of symbols for the classification of patents according to the different areas of technology to which they pertain.[3] A generic IPC category consists

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

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

Table 5.1: International Patent Classification (IPC) sections and their description. The last column reports the number of patents registered in our data set for the corresponding IPC section.

of a letter, the so-called "section symbol", followed by two digits, the so-called "class symbol", and a final letter, the "subclass". This four-character term is then followed by a group/subgroup indication, represented by additional digits. A typical IPC term can be written as follows: B34H 6/99. The sections identified by the IPC are historically stable and amount to 8, from A (human necessities) to H (electricity). The lower levels are instead subject to more frequent revisions; the eighth and last IPC edition consists of more than 120 classes, 600 subclasses, 7'000 main groups and 60'000 subgroups.

The titles of the 8 sections, as well as a patent count for each section in our data set, is reported in Table 5.1. We find that the number of patents in all sections reflects their technological dynamism [187]. Indeed, all sections are not equally represented. For example, the two sections with the lowest patent counts are Textiles, Paper and Fixed Constructions (less than 20 000 patents), two typical mature industries, while the sections of Physics and Electricity has the highest patent count (about 50 000 patents). In these sections, patents are often filed by firms belonging to industrial sectors where products innovation and radical innovations play a major role (e.g., from firms working on computer hardware, computer software and electronic components).

In contrast to the IPC, the ISI-OST-INPI classification scheme is more adapted to the *technological* knowledge space for patents data [200]. As suggested above, this scheme was developed by the Fraunhofer ISI, the Observatoire des Sciences et des Techniques (OST) and French patent office (INPI) in order to overcome problems in the IPC and the US classification scheme. There

exist various versions of ISI-OST-INPI classification and we chose to use the most updated one, available from PATSTAT, Patent Statistical Database[4]. In this version, the scheme groups different IPC codes into 5 technology areas, which are again divided in a total of 35 fields. The main 5 areas are: 1) Electrical engineering 2) Instruments, 3) Chemistry, 4) Mechanical engineering and 5) Other fields. In table 5.2, we report as an example the classification scheme for the technology area `Electrical engineering`, as provided from table `tls901_techn_field_ipc` available in PATSTAT Online, edition Autumn 2016. In each entry of the table there is an ISI-OST-INPI code with the corresponding name of the field and IPC codes. We have created similar tables also for the other four technology areas (not shown). Using these tables, we assigned to the patents present in our database with one or more IPC codes new ISI-OST-INPI codes. Our matching procedure was successful since it worked for about 99% of the patents.

We intend to test our model on a broad set of firms, belonging to several industrial sectors, and therefore exhibiting patent activities distributed across all sections, classes and subclasses. For this reason, we have only considered the 8 dimensions (i.e. the first letter) of the IPC code, and the 35 dimensions of the ISI-OST-INPI code. Choosing a more refined class- or subclass-level division would result in an excessive patent granularity, meaning a even higher dimensionality for the corresponding knowledge space. However, comparing the results for the 8- and the 35-dimensional knowledge space already allows us to draw conclusions about the robustness of our findings with respect to the dimensionality of the knowledge space.

Knowledge position To ensure a match with our model representation, we define the knowledge position of a firm $\mathbf{x}_i \equiv (x_{i1}, x_{i2}, \ldots, x_{iD})$ as the set of normalized patent counts x_{is} in each class $s = 1, 2, \ldots D$ (where D is the maximum number of dimensions in the respective classification scheme, i.e. either 8 or 35):

$$x_{is} \equiv \frac{N_{is}}{\sum_s N_{is}} \qquad s = 1, \ldots, D \tag{5.1}$$

[4] `https://www.epo.org/searching-for-patents/business/patstat.html`

	Electrical engineering	
1	Electrical machinery, apparatus, energy	F21H, F21K, F21L, F21S, F21V, F21W, F21Y, H01B, H01C, H01F, H01G, H01H, H01J, H01K, H01M, H01R, H01T, H02B, H02G, H02H, H02J, H02K, H02M, H02N, H02P, H02S, H05B, H05C, H05F, H99Z
2	Audio-visual technology	G09F, G09G, G11B, H04N 3, H04N 5, H04N 7, H04N 9, H04N 11, H04N 13, H04N 15, H04N 17, H04N 19, H04N 101, H04R, H04S, H05K
3	Telecommunications	G08C, H01P, H01Q, H04B, H04H, H04J, H04K, H04M, H04N 1, H04Q
4	Digital communication	H04L, H04N 21, H04W
5	Basic communication processes	H03B, H03C, H03D, H03F, H03G, H03H, H03J, H03K, H03L, H03M
6	Computer technology	G06C, G06D, G06E, G06F, G06G, G06J, G06K, G06M, G06N, G06T, G10L, G11C
7	IT methods for management	G06Q
8	Semiconductors	H01L

Table 5.2: ISI-OST-INPI classification scheme based on the IPC, for the technology area of Electrical engineering. The first column is the ISI-OST-INPI code, the second gives the name of the field and the third column groups the different IPC codes corresponding to the same ISI-OST-INPI code.

N_{is} is the number of patents that the firm i has in a given class s. In order to compute knowledge distances between pairs of firms, we use the Euclidean metric, similar to Tomasello *et al.* [226]. This means that the knowledge distance between two firms i and j reads as:

$$|\mathbf{x}_i - \mathbf{x}_j| = \sqrt{\sum_{s=1}^{D} (x_{is} - x_{js})^2} \tag{5.2}$$

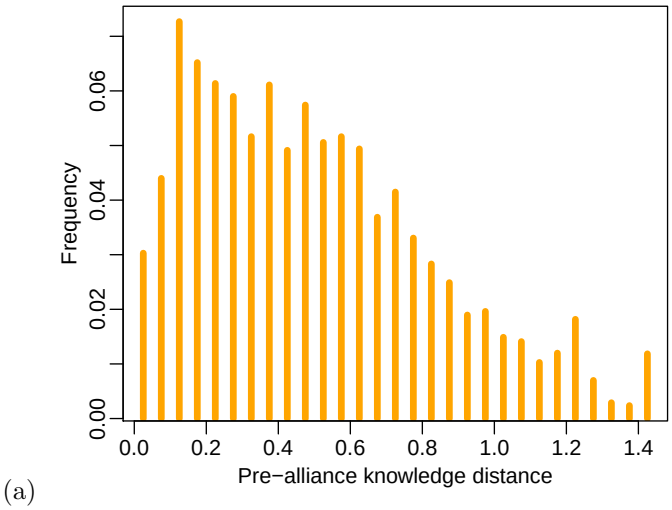
In Figs.5.1(a,b) we provide a visualization of the knowledge positions of firms using the two patent classification schemes. In the time-aggregated R&D network, nodes represent firms and their colors depend on the patents they have filed between 1984 and 2009. In Fig.5.1(a), we have assigned different to each firm the color of IPC-section where it has filed more patents. With this, we

approximate the knowledge position of each firm for visualization purposes. In Fig.5.1(b), we apply the same procedure but considering the 5 main areas of the ISI-OST-INPI classification scheme. From both figures, we find that the two main clusters, which are comprised mainly by pharmaceutical companies (bottom cluster) and firms working on computer hardware, software and communications (top cluster), are dominated by few colors. This shows that most alliances occur among firms with a similar knowledge base; alliances with different knowledge bases occur only in specific combinations.

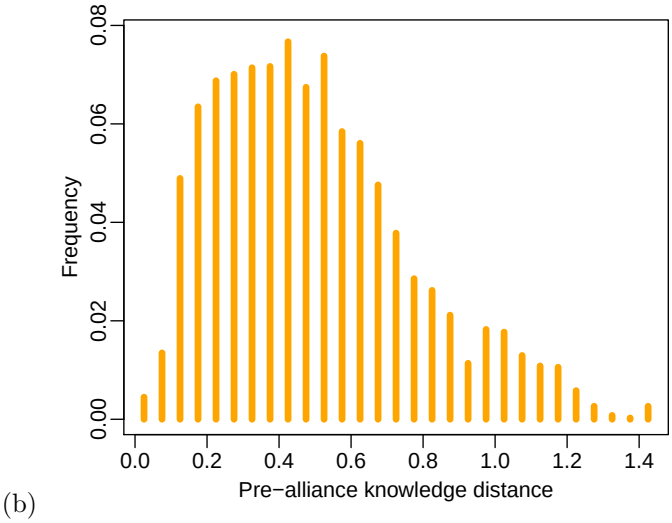
Distributions of pre-alliance knowledge distances Using the definitions provided in Eqs. (5.1) and (5.2), we can now compute the knowledge positions of the 5'168 firms listed in our data set for the two different classification schemes together with the knowledge position of their alliance partners. This allows us to calculate the distribution of the knowledge distances between every pair of allied firms, at the moment of alliance formation (which we know precisely). We save these pre-alliance distances together with the positions of the firms in knowledge space, to later use this information for setting up the computer simulations.

In Fig. 5.2 we report the distributions of *pre-alliance knowledge distances* for the two different classification schemes. The minimum observed value of knowledge distance is 0, while the maximum equals $\sqrt{2}$ (see Eq.(5.2)), for normalization reasons. We find, for both schemes, that the distribution is peaked around an intermediate distance and left-skewed, i.e. shifted toward small values. Interestingly, we observe that the counts drop when such distances approach zero, meaning that firms with the exact same patenting activity tend not to form alliances. In addition, it is important to remark that the granularity of the different schemes does not impact the distributions.

When computing the empirical knowledge position $\mathbf{x}_i$ of a firm at a given date t, we consider all the patents for which the firm has applied in a preceding time window $[t-\Delta t, t]$. To have a reliable and updated measurement, without losing at the same time too much patent information due to a short time window, we use $\Delta t = 5$ years. We have tested different time windows, ranging from 1 to 10 years, and have found that this only increases the number of missing observations or the noise in the distributions, with no effect on our results.



(a)



(b)

Figure 5.2: Empirical knowledge distance between every pair of partnered firms, as of the day preceding the alliance formation, calculated in (a) the 8 dimensional knowledge space defined by the IPC scheme and in (b) the 35 dimensional knowledge space defined by the ISI-OST-INPI classification scheme.

5.3 The model

We now describe the microscopic interaction rules of our agent-based model. In a first phase, the agents form links based on their network features and their social capital; we call this the "exploration (link formation)" phase. Subsequently, they exchange knowledge through these links, thus approaching each other in a metric knowledge space; we call this "exploitation (knowledge transfer)" phase. While exchanging knowledge, agents can also form new links; in addition, each link can be terminated with a given probability. Hence, the exploration and exploitation phases are not separated in time.

5.3.1 Exploration: link formation

Activation. We consider a network composed of N agents. Each agent represents an agent that is endowed with two fundamental attributes, an *activity* and a *label*. The activity a_i of agent i defines her propensity to engage in a collaboration event. We obtain a_i from the distribution of empirical activities extracted from the SDC alliance data set (see Section 5.2). At every time step, agent i initiates an alliance with probability $p_i = \eta a_i \mathrm{d}t$. Consequently, the number of active agents per time step is $N_A = \eta \langle a \rangle N \mathrm{d}t$. Here $\langle a \rangle$ is the average agent activity and η is a rescaling factor that allows to adjust the activation rates. We fix $\eta = 0.0115$ to obtain N_A close to 2 which is the number of active firms per day reported in the alliance data set. More details will follow on the interpretation of the time step duration $\mathrm{d}t$.

Alliance size. Upon activation, an agent selects the number of partners for a collaboration. We simulate this selection by sampling without replacement a value n from the empirical distribution of alliance sizes, directly measured from the SDC Platinum data set. With this, we assume that the number of partners, $m = n - 1$, with whom the alliance is formed is independent of any characteristic of the active agent.

Label propagation. The second key attribute, called *label*, is used to model the belonging of firms to communities that are implicitly defined through shared practices and/or behaviors. In other words, a label can be thought as a membership to a well defined and recognized "club" or "circle of influence". We assume that such membership is unique and fixed, i.e. an agent cannot change it nor have more than one. At the beginning of each simulation all agents are non-labeled. They can obtain a label in two different ways, (i) by being selected as partner for an alliance or (ii) by initiating one. In the former case, the non-labeled agent receives the label of the initiator of the alliance, while in the latter she receives a new label that no other agent has in the network. Both cases are illustrated in Fig. 5.3. It was shown that the described label propagation mechanism can very effectively explain the presence of clusters, or communities, in R&D networks [224].

Selection of the partner categories. The presence of labels allows to distinguish between different types of alliances, dependent on the initiator. If the initiator is a labeled agent, she can link to an agent with the same label (with probability p_s^L), to an agent with a different label (p_d^L), or to an agent without a label (p_n^L). If the initiator is a non-labeled agent, i.e. she is a *newcomer* in the collaboration network, she can link to a labeled agent (with probability $p_l^{\mathcal{N}}$), or to another non-labeled agent ($p_n^{\mathcal{N}}$). The link formation with a labeled agent (described by the probabilities p_s^L, p_d^L and $p_l^{\mathcal{N}}$) describes *endogenous mechanisms*, because the initiator of the alliance has information about the network position (i.e. social capital) of its potential partners. For this case, the two linking probabilities p_s^L and p_d^L allow to tune the importance of the *cohesiveness* as an endogenous driver. The connection with a non-labeled agent (events p_n^L and $p_n^{\mathcal{N}}$) describes *exogenous mechanisms* because, in this case, the initiator has no information about the social capital of an agent that is not yet part of the network.

Link formation. Once the category (label) of each partner is determined, the initiator of the alliance selects the specific partner. To do this, we employ a linear preferential attachment rule, where a agent j is selected with probability proportional to her degree k_j (i.e., the number of previous collaborations

with distinct partners). This rule is chosen to capture the *prominence* of a firm, namely the history of its previous alliances, as an endogenous driver. Obviously, this does not apply when the initiator, labeled or not, decides to connect to a non-labeled agent, which has by definition no previous partners ($k_j = 0$). In this case, the partner is selected among all non-labeled agents with equal probability. When the selection process is complete, the initiator connects to its m partners, which accept the offer. A variant of the model in which partners can also reject the offer is discussed in [222]. In agreement with our representation of the R&D network, we assume that all the m partners will also link to each other, forming a fully connected clique of size $n = m + 1$ with $m(m + 1)/2$ links (see Fig. 5.3).

5.3.2 Exploitation: knowledge transfer

The second set of microscopic rules models the process of knowledge exchange between pairs of collaborating agents, similar to what has been investigated in Tomasello *et al.* [226]. Basically, we assume that every agent in the network is located in a metric knowledge space and, as a consequence of its collaborations, approaches its partners in this space. In case of multiple partners, the motion of the focal agent is determined by the vectorial sum of the effects of all of its partners.

Location in a metric knowledge space. Here we refer to the description of the (two different) knowledge spaces given in Sect. 5.2.2. Every agent i ($i = 1, \ldots, n$) is characterized by a D-dimensional vector $\mathbf{x}_i \equiv (x_{i1}, x_{i2}, \ldots, x_{iD})$, where the components $x_{i1}, x_{i2}, \ldots$ are real numbers ranging from 0 to 1. In the case of R&D networks, these numbers are given by the ratios of patents, reflecting the firm's expertise in each of the D dimensions. Only $D - 1$ values of the x_{is} are independent because of the boundary condition that the patent fractions have to sum up to 1. The dimension of the knowledge space, D, is a structural characteristic of the system and fixed depending on the classification scheme and granularity selected to classify the patents.

Approaching in the metric knowledge space. We assume that the existence of a link causes the involved agents to exchange knowledge with their partners and to align their knowledge bases. Hence, as a result of this exchange, they should approach each other in knowledge space. To capture this dynamics, every agent is characterized by a *learning rate* μ. This parameter is, in first approximation, constant over time and the same for all agents in the collaboration network. The model dynamics equation can be written as follows:

$$\dot{\mathbf{x}}_i(t) = \mu \sum_{j \in \mathcal{N}_i(t)} [\mathbf{x}_j(t) - \mathbf{x}_i(t)] \tag{5.3}$$

where $\mathcal{N}_i(t)$ is the set of partners of the agent i at time t. For implementing the model in computer simulations, we use discrete time steps of length $\mathrm{d}t$. The evolution of every agent's position $\mathbf{x_i}$ can then be expressed as:

$$\mathbf{x}_i(t + \mathrm{d}t) = \mathbf{x}_i(t) + \mu \sum_{j \in \mathcal{N}_i(t)} [\mathbf{x}_j(t) - \mathbf{x}_i(t)]\, \mathrm{d}t \tag{5.4}$$

It should be noted from Eq. (5.3) that the speed at which a collaborating agent moves in the knowledge space is given by the product of two factors: μ – the approach *rate* – and its distance from the partners. With this dynamics, the farther agents are in the knowledge space, the faster they move towards each other. When the agents' distance decrease, the potential for new learning from the collaboration and consequently the approaching speed decrease as well. This, eventually, motivates to cancel the collaboration and to terminate the alliance after some time.

Although the dynamics of knowledge exchange is quite simple, it has a number of implications we would like to point out. First of all, in the present model proximity in knowledge space is not a precondition for the agents' interactions. This is different from other existing models [see, for instance, 16, 80, 226] where some sort of "similarity" is assumed for a possible collaboration. In our model collaboration is determined by the network formation mechanisms, where the different link probabilities are independent of the agents' knowledge positions.

Second, in our model every link (i.e. every collaboration) necessarily implies that the involved partners approach each other in the knowledge space. This

Parameter	Explanation	
$p_s^{\mathcal{L}}$	Prob. for a ***L*abeled** agent to chose an agent with ***s*ame label**	(NF)
$p_d^{\mathcal{L}}$	Prob. for a ***L*abeled** agent to chose an agent with ***d*ifferent label**	(NF)
$p_n^{\mathcal{N}}$	Prob. for a ***N*on-labeled** agent to chose a ***n*on-labeled** agent	(NF)
μ	Approaching rate in the knowledge space	(KE)
τ	Link characteristic life time	(KE)

Table 5.3: Model parameters and their description. The "network formation" (NF) parameters are associated with the creation of new links in the collaboration network. The "knowledge exchange" (KE) parameters are associated with the approach of the agents in a metric knowledge space, occurring as a consequence of a collaboration.

reflects the purpose of the network formation, namely exchange of knowledge. Our dynamics assumes that agents approach each other in *all* dimensions of the knowledge space, not just in one particular dimension representing their area of collaboration. This reflects the effect of *knowledge spillovers* [16], i.e. agents profit from the collaboration not just by the exchange of specific knowledge, but also by learning more general experience.

Alliance termination. R&D alliances have been proven to have a finite duration [163, 223]. In order to develop a realistic model, we introduce as a key parameter the characteristic life time τ of a link. Assuming that the durations of alliances are distributed according to a Poisson process with rate $1/\tau$, the *mean duration* is obviously equal to τ. In our computer simulations, which use discrete time steps of length $\mathrm{d}t$, this translates into the use of a fixed termination probability $p_{\mathrm{T}} = \mathrm{d}t/\tau$ for any link at any time step.

To keep a simplistic set of rules, we assume that the parameter τ is a constant, independent of any other feature of the network or the knowledge exchange dynamics or the knowledge stock of the agents. One possible extension would be to link τ to the knowledge distance of the two partners, or some other network-related feature.

To sum up, in this section we have described a set of microscopic rules which aim at reproducing the formation of links in a collaboration network, together with the approach of the agents in an underlying knowledge space. We summarize the model microscopic rules by means of a visual example in Fig. 5.3 and report the nomenclature of all parameters in Table 5.3.

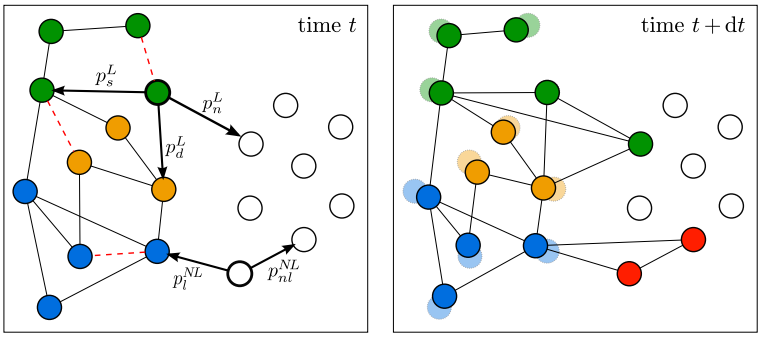


Figure 5.3: A representative example of network evolution in a bi-dimensional ($D = 2$) knowledge space. The position of the agents in the plot corresponds to their coordinates in the knowledge space. At time $t + \mathrm{d}t$, all existing links cause the respective agents to approach in the knowledge space. Furthermore, we illustrate two collaboration events occurring at time t. The first one is initiated by a labeled agent (in green), that has linked to $m = 3$ new partners, forming a fully connected clique. The second one is initiated by a non-labeled agent, that has linked to $m = 2$ new partners and has taken a new arbitrary label (red). At time $t + \mathrm{d}t$, the alliance initiators propagate their labels (respectively, the green one and the red one) to the partners that were not labeled at time t yet. Finally, we illustrate the termination of 3 links (depicted with red dashed lines) at time t.

5.4 Model calibration: a two-step procedure

We now calibrate our model against the data, to estimate the value of its parameters. This is performed in two steps, for network formation and knowledge exchange, by using two data sets, R&D alliances and patents.

5.4.1 Network formation parameters

In the *first step*, calibrating the network formation model, we fix a set of parameters that we can directly measure from the data, namely the number of agents N =5'168, the distribution of the agents activities a_i, and the distribution of number of partners m per alliance event.

We then estimate the remaining parameters, i.e. p_s^L, p_d^L and $p_n^{\mathcal{N}}$, by running a set of computer simulations, to identify the simulated collaboration network that matches best with the alliance data set. We stop every computer simulation when the total number of formed alliances equals the number of alliance events reported in the SDC data set, E =7'417. We vary the values of p_s^L, p_d^L and $p_n^{\mathcal{N}}$ in discrete steps spaced by 0.05, in the interval $(0, 1)$. The parameters p_s^L and p_d^L are bounded by the condition $p_n^L = 1 - p_s^L - p_d^L \geq 0$, meaning that their sum has to be smaller or equal to 1. This condition translates into 3'249 points to explore in the 3-dimensional parameter space, for each of which we run 100 simulations (for a total of 324'900 runs).

The networks that we generate by means of computer simulations are matched to the data with respect to three global indicators: average degree $\langle k \rangle$, average path length $\langle l \rangle$, and global clustering coefficient C [5]. For the empirically observed R&D network, we denote such measures as $\langle k \rangle^{\text{obs}}$, $\langle l \rangle^{\text{obs}}$, and C^{obs}, respectively, and their values are $\langle k \rangle^{\text{obs}} = 3.45$, $\langle l \rangle^{\text{obs}} = 5.05$ and $C^{\text{obs}} = 0.11$.

In order to identify which parameter combination is able to give the best match with the real R&D network, we use a Maximum Likelihood approach, similar to Tomasello *et al.* [224]. We do not have a set of observations against which we can calibrate our model; instead, we only have one empirical point: the real R&D network. In particular, we cannot consider the three measures $\langle k \rangle$, $\langle l \rangle$ and C as independent, therefore the Likelihood function $\mathcal{L}$ reads as:

$$\mathcal{L}(p|V^{\text{obs}}) = f(V^{\text{obs}}|p) \tag{5.5}$$

where $f(\cdot)$ is the joint density function of all parameter combinations p resulting in a network that is equivalent to the observed one, G^{obs}. Both p and V^{obs} are vectors with three components, expressing respectively the three model parameters $p \equiv (p_s^L, p_d^L, p_n^{\mathcal{N}})$ and the three global network measures $V^{\text{obs}} \equiv \left(\langle k \rangle^{\text{obs}}, \langle l \rangle^{\text{obs}}, C^{\text{obs}}\right)$. Therefore, we need to find the parameter combination $(p_s^L, p_d^L, p_n^{\mathcal{N}})$ maximizing the Likelihood $\mathcal{L}(p|V^{\text{obs}})$ to generate a network whose macroscopic properties are *sufficiently similar* to the real network G^{obs}. By this, we mean that the relative errors from the observed values for

[5] For a rigorous definition of these measures, see Tomasello *et al.* [224].

Optimal simulated R&D network				**Real R&D network**[6]	
Model parameter	Value	Measure	Value	Measure	Value
p_s^{*L}	0.45	$\langle k \rangle^*$	3.48 ± 0.01	$\langle k \rangle^{\text{obs}}$	3.45
p_d^{*L}	0.2	$\langle l \rangle^*$	5.02 ± 0.08	$\langle l \rangle^{\text{obs}}$	5.05
p_n^{*L}	0.35	C^*	0.111 ± 0.007	C^{obs}	0.109
$p_n^{*\mathcal{N}}$	0.1				
$p_l^{*\mathcal{N}}$	0.9				

Table 5.4: Link formation parameters p^* defining the optimal simulated R&D network. The average degree, average path length and global clustering coefficient of the 100 realizations of the optimal R&D network are compared to their empirical counterparts.

the average degree $\varepsilon_{\langle k \rangle}$, the average path length $\varepsilon_{\langle l \rangle}$ and the global clustering coefficient ε_C have to be smaller than a certain threshold ε^0.

We empirically compute the Likelihood function $\mathcal{L}$ for each point in the parameter space by counting the fraction of its 100 simulation realizations that fulfill the criteria $\varepsilon_{\langle k \rangle} < \varepsilon^0$; $\varepsilon_{\langle l \rangle} < \varepsilon^0$; $\varepsilon_C < \varepsilon^0$. This way, we obtain values that can range from 0 (no realization of that parameter combination fulfills the criteria) to 1 (all of its realizations fulfill the criteria). For the choice of the error threshold ε^0, we take a conservative approach and use $\varepsilon^0 = 0.02$, that ensures a good matching with the real R&D network, without cutting out too many points in the parameter space.

We find that the point in the parameter space with the highest likelihood score has coordinates: $p_s^{*L} = 0.45$, $p_d^{*L} = 0.2$ and $p_n^{*\mathcal{N}} = 0.1$. This means that labeled agents show a fairly balanced alliance strategy, with $p_s^{*L} = 0.45$, $p_d^{*L} = 0.2$, and consequently $p_n^{*L} = 0.35$, while non-labeled agents connect rarely with other non-labeled agents ($p_n^{*\mathcal{N}} = 0.1$) and prefer to link with labeled ones ($p_l^{*\mathcal{N}} = 0.9$). In Table 5.4, we report the full set of parameter values maximizing the likelihood score, together with the values of average degree, average path length and global clustering coefficient for the simulated and the real R&D networks.

These results are in line with those presented by Tomasello *et al.* [224]. However, the R&D network with patent data, used here, exhibits an even stronger tendency to favor connections with labeled agents (i.e. incumbent firms) than

the pooled R&D network including all firms, irrespectively of their patenting activity. Let us spend a few words on the comparison between these two networks.

Due to the fact that our analysis in now restricted only to firms for which patent data are available, one could expect either an increase in the importance of network endogenous mechanisms, given that we are considering, on the one hand, larger and more active firms – or an increase in the importance of exogenous mechanisms, given that we are considering, on the other hand, firms for which the technological dimension could be more relevant in the alliance formation strategy. Our data confirm the first hypothesis, that is the increase in the relevance of network endogenous mechanisms, which results in higher probabilities for the agents to collaborate with agents that are already part of the network, and therefore already labeled. This behavior is present irrespective of whether the alliance event is initiated by a labeled or a non-labeled agent: precisely, 65% of the collaborations initiated by labeled agents ($p_s^{*L} + p_d^{*L}$), as well as 90% of the collaborations initiated by non-labeled agents ($p_l^{*\mathcal{N}}$), involve a labeled agent as a partner.

5.4.2 Knowledge exchange parameters

In the *second step*, we fix the network formation parameters to the values obtained in the first step, and run a second set of computer simulations. This time we estimate the knowledge exchange parameters, i.e. μ and τ, by identifying the simulated collaboration network that best matches with the patent data set. To quantify the knowledge space, we use either the eight main sections of the IPC scheme or the 35 technological fields of the ISI-OST-INPI classification scheme, i.e. the dimensions are set to $D = 8$ or $D = 35$.

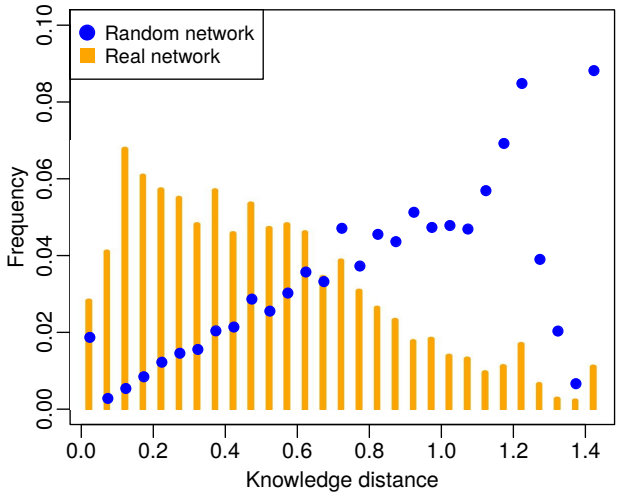
Pre-alliance conditions In order to calibrate the dynamics of knowledge transfer, we need to assign to the agents a current position in the respective knowledge space, to calculate their future positions. Following the model of network formation, we need to distinguish between the agent that initiates a collaboration (when becoming active), and the m collaborators chosen by the initiator.

A naive approach would assume that we first randomly choose an initiator with its initial position in knowledge space, then randomly choose m collaborators, their distances in knowledge space randomly sampled from the empirical distribution of pre-alliance distances shown in Figure 5.2. Second, we run the knowledge exchange dynamics of Eq. (5.2), to calculate the expected movement in knowledge space for a given set of parameters τ, μ. Eventually, we compare the distribution of distances for various τ, μ with the empirical distribution of post-alliance knowledge distances, to find out which set of parameters matches best.

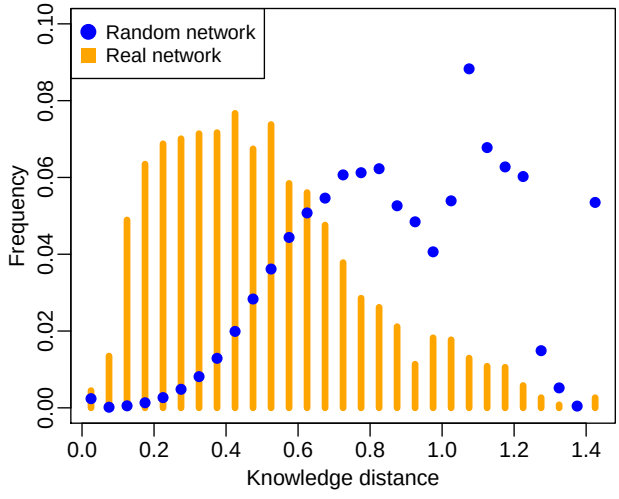
While the second part of the procedure is correct, the first part is based on the wrong assumption that firms *randomly* choose their collaboration partners from the knowledge space. Figure 5.4 shows, for the two different knowledge metrics used, how the distribution of pre-alliance distances should look like if every possible knowledge distance would be realized. We note the strong deviations between the random and the empirical distributions. First, the random distributions appear right-skewed while the empirical are left-skewed. Second, the average pre-alliance distance are around 0.9 in the random case, while the averages of the empirical pre-alliance distances is much smaller, around 0.6.

With this, we can conclude that the empirical pre-alliance distance distributions cannot be explained by assuming that firms create alliances without considering the position of their possible collaborators in the knowledge space. Hence, we need to essentially consider the *full* agent-based model – not to calibrate the dynamics of knowledge exchange, but to correctly determine the *initial conditions* for the knowledge exchange dynamics. This lends strong support to consider the *combined processes* of network formation and knowledge exchange, as it is proposed in our model, instead of investigating knowledge exchange in isolation.

In order to determine the pre-alliance conditions in knowledge space for our model at a given time t, we distinguish between agents that are *not* currently, at time t, involved in any collaboration and those that *are* currently involved. Agents that *are* involved, already have a position in knowledge space that reflects their previous interaction with other agents during the simulation up



(a)



(b)

Figure 5.4: Pre-alliance distance distributions from the empirical and a randomized R&D network. In (a) we used the IPC scheme to calculate the firms positions, while in (b) the ISI-OST-INPI scheme.

to time t. Thus, we decide to keep these (simulated) positions at time t as starting point for their knowledge exchange in the new alliance. For those agents that are *not* involved in a collaboration at time t, we obtain the initial conditions from sampling from the empirical data. Precisely, the position of an initiator that is not currently involved in an alliance is sampled from the distribution of pre-alliance positions obtained from the real patent data. And for the collaborating agents that are *not* involved in any other alliance at time t, we assign a *knowledge distance* by sampling with replacement from the empirical distribution of pre-alliance distances given in Figure 5.2.

This procedure of determining the pre-alliance distance distribution mixes up two conceptually different information. Part of it is obtained from *simulations*, this way taking into account the path dependence of the recent history in collaborations, i.e. the active partners in alliances and their influence on knowledge exchange. Another part of information comes from the *empirical distribution* of pre-alliance knowledge positions/distances that reflects e.g. preferences of agents in choosing partners at shorter distances. Further, it captures the fact that firms not engaged in any R&D alliance can still perform related activities and thus move in knowledge space, which is reflected by their new position assigned when engaging in a new alliance. We emphasize again that, without the empirical information, we would randomly pair agents that likely had not chosen to collaborate or we would assume that agents do not move in knowledge space by themselves. Without the simulations, on the other hand, we would create problematic artifacts in all cases where agents already involved in a collaboration are chosen to participate in a new alliance. In such cases, we cannot assign two positions in knowledge space to the same agent or randomly switch between profiles. Thus, the best solution is to keep the evolution of agents *during* existing collaborations into account, as a precondition for new ones.

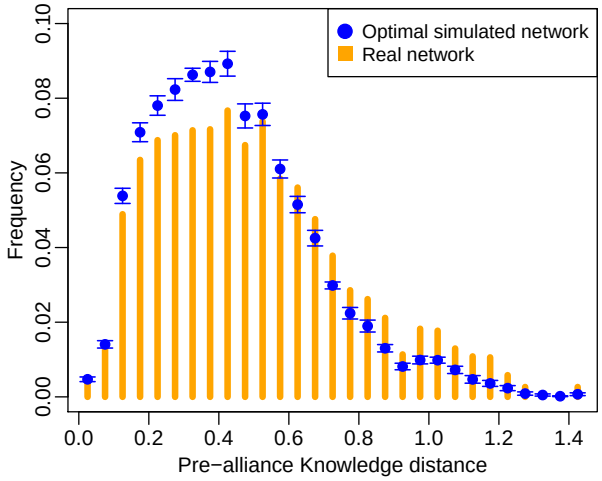
This leads us to an important question that we need to answer before we can discuss the details of the parameter calibration: What is the error that we may introduce by mixing these two source of information for determining the initial conditions? In Fig. 5.5(a), we show the distribution of pre-alliance distances that follows from the constraint of respecting current knowledge positions in comparison to the empirical distribution. We find that the simulated distri-

bution matches the empirical one over a large range; however, the simulations overestimate the probability of having alliances among firms separated by a small knowledge distance. This deviation is significant only in the range of distances between 0.2 and 0.4, where the distribution has its maximum.

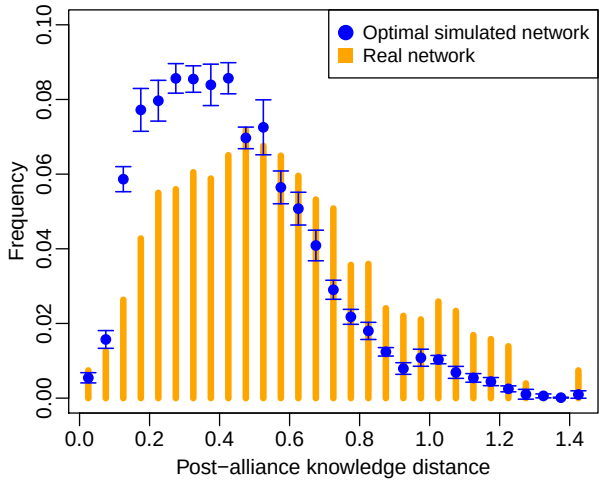
Obviously, such deviations in the initial conditions are amplified during the simulated knowledge exchange, as can be seen in Fig. 5.5(b) which shows the *post-alliance distance distribution*. Precisely, compared to the empirical distribution of *pre*-alliance distances, in the empirical distribution of *post*-alliance distances the probability to have a small knowledge distances has decreased, whereas it has increased in the corresponding simulations. We will comment on this interesting observation further in Sect. 5.6.

At this point, we just emphasize that the empirical distribution of pre-alliance distances is much better matched by the distribution obtained from our simulations that use the selection process described above (see Fig. 5.5(a)) compared to the distribution obtained assuming a random selection process (see Fig. 5.4(b)). Indeed, when we perform a two-sided Kolmogorov-Smirnov (KS) test between our simulated distribution of pre-alliance distances and the empirical one, we find an average $\mathcal{D}$-statistic 10 times smaller, i.e. better, compared to the $\mathcal{D}$-statistic coming from the KS-test performed between the distributions shown in Fig. 5.4(b). We disregard the p-value of the KS-test, because we are not interested in statistically inferring the provenience of the two distributions from a hypothetical common distribution. Our aim is instead to quantify the similarity between pairs of distributions, a measure that is already fully captured by the $\mathcal{D}$-statistics of a two-sided KS-test. Hence, in the following we will take the distribution of pre-alliance distances shown in Fig. 5.5(a) as a good proxy for the initial condition at the moment of alliance formation.

Optimal parameters In the subsequent computer simulations we vary the values of the two remaining knowledge exchange parameters, i.e. the agents' approaching rate μ and the characteristic alliance life time τ. We consider the values 5×10^{-8}, 10^{-7}, 5×10^{-7}, 10^{-6}, 5×10^{-6}, 10^{-5} for the parameter μ and the values 700, 1000, 1500 and 2000 for the parameter τ, thus having a total of 24 points to explore in the parameter space. The interpretation of the parameter τ is straightforward: as explained in Section 5.3.1, we adjust



(a)



(b)

Figure 5.5: Empirical and simulated distances between firms at the moment of alliance formation and at the assumed termination of alliances after $\tau = 700$ days. In both plots the distances are calculated in the 35 dimensional space defined by the ISI-OST-INPI classification, the blue circles correspond to the mean values and the error bars correspond to the standard deviations of all the measures we study on the 100 realizations of the optimal simulated R&D network.

the activation rate of the agents such that the length of a time step dt can be directly interpreted as 1 day. Therefore, the value of τ, which is by design expressed in time steps, can be thought of as the characteristic duration of a real alliance in days.

For each of the 24 parameter combinations, we run 100 simulations that combine the network formation process (using the optimal parameters determined) and the knowledge exchange dynamics. This results in a total of 2'400 simulation runs only to complete the second step of our calibration procedure, namely to determine the optimal knowledge exchange parameters. We store the distributions of post-alliance knowledge distances and knowledge distance shifts in each run. Similar to the first step, we stop every computer simulation when the total number of collaborations equals the number of alliance events reported in the SDC data set, E =7'417.

As explained, the distribution of pre-alliance distances shown in Fig. 5.5(a) is used as an input of the simulations. Thus, we use the distribution of post-alliance knowledge distances, obtained from each of the 100 simulations for every parameter combination, to compare it to the respective distance distribution obtained from the empirical R&D network. This comparison relies on determining the *post-alliance* time. It becomes a problem for the empirical data because the termination dates of alliances are not available. In the simulations, however, we have assumed that alliances have a duration τ and are terminated stochastically, afterwards. To allow for comparison, we compute, from the empirical data, the knowledge distance between every pair of linked firms after the same time period τ, in days, as used in the corresponding simulation.

To compare the two distributions of simulated and empirical knowledge distances, we use the two-sided KS-test that assigns a score, the $\mathcal{D}-$statistics, to each simulated distribution. The value of the $\mathcal{D}-$statistics decreases as the simulated and the empirical distributions become more similar, hence, it is used here as goodness score for each simulation. We finally average the 100 score values for the 100 simulations, for each combination of the parameters.

The resulting goodness scores are presented in the heat map plot of Fig. 5.6. It shows the bi-dimensional parameter space of alliance duration τ and learning

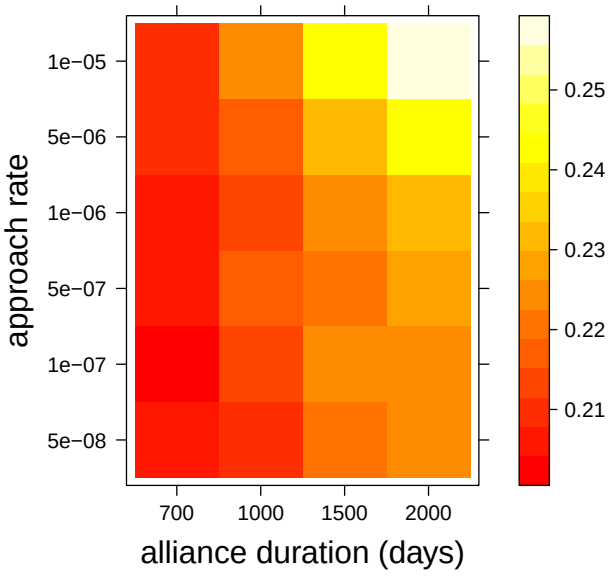


Figure 5.6: Goodness score for every point in the parameter space, depicted by means of a heat-map. The color scale corresponds to the score value; the lower the score, the closer the simulated distribution of post-alliance distances is to the empirical one. The simulations and the distances have been obtained considering the 35 dimensional space defined by the ISI-OST-INPI classification scheme.

rate μ. As the color code indicates, we find an entire region of parameters with maximized goodness score for parameter combinations with medium to large μ, but low τ values.

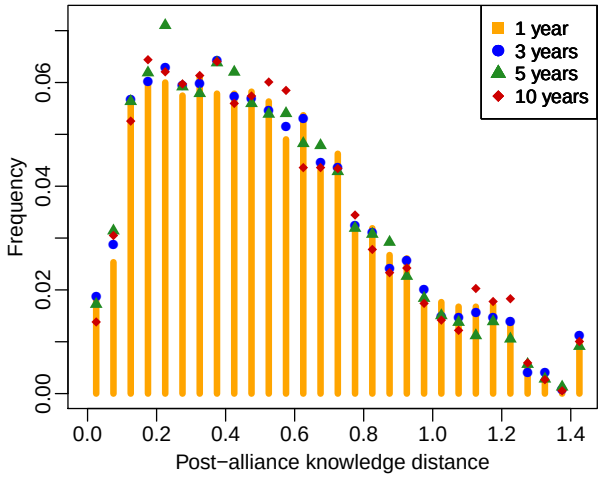
Although many parameter combinations exhibit a similar, low goodness score, i.e. they are fairly equally able to reproduce the empirical post-alliance knowledge distance distribution, the best parameter sets can be ranked quantitatively. We find that the parameter point yielding the best goodness score is identified by the following coordinates: $\mu = 10^{-7}$ and $\tau = 700$. This means the optimal simulated collaboration network exhibits a low approaching rate, and

a characteristic alliance duration slightly shorter than 2 years. This is not only consistent with previous theoretical and empirical observations [99, 163], but also in line with our previous assumption Tomasello *et al.* [223] to terminate alliances after 3 years in the empirical network representation. Taking into account that we have obtained this result here by using two different data sets and an involved agent-based model, the agreement is even more remarkable.

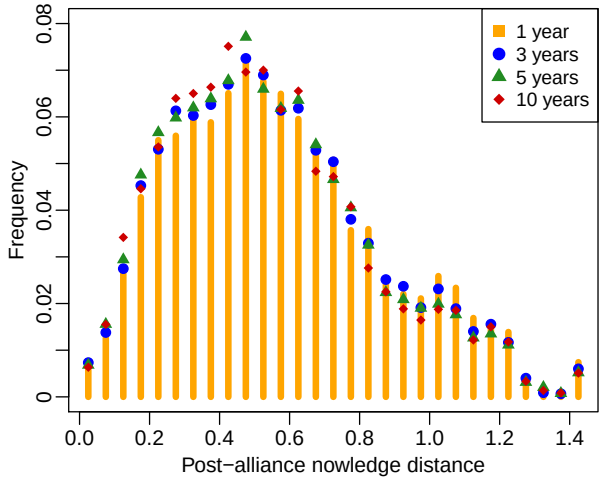
5.4.3 Robustness analysis

Distribution of post-alliance knowledge distances. Already for the model calibration, we addressed the problem that the exact durations of R&D alliances are not known from the data set. Hence, the above estimations of the optimal duration τ is conditional on the knowledge transfer rate μ. However, we can also independently investigate how sensitive the distribution of post-alliance distances responds to changes of the (unknown) duration of alliances. This is done in the following two steps for both of the knowledge space metrics used.

In the first step, we analyze the *empirical distribution of knowledge distances* for different alliance durations. The NBER patent data set has a time granularity of 1 year. This forces us to use time increments of 1 year with a minimum window of 1 year. In Fig. 5.7 we show the post-alliance knowledge distance distribution for different time windows: 1, 3, 5 and 10 years. We find that, for both knowledge space metrics, the shape of the knowledge distance distribution appears to have the same shape, irrespective of the time window chosen. This allows for two conclusions. First, an assumed increase of the alliance duration does not considerably impact the post-alliance distance distribution, most likely because firms do not move much in knowledge space over time. Second, because of this our modeling approach is robust against the (unknown) duration of alliances. There is a firm relation between τ and μ as discussed in Figure 5.6. But even for larger durations τ, the properly calibrated model can be used to reproduce the empirical distribution of post-alliance knowledge distances.

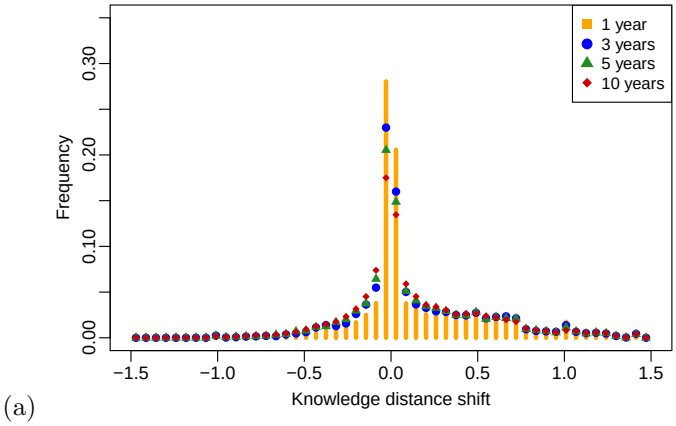


(a)

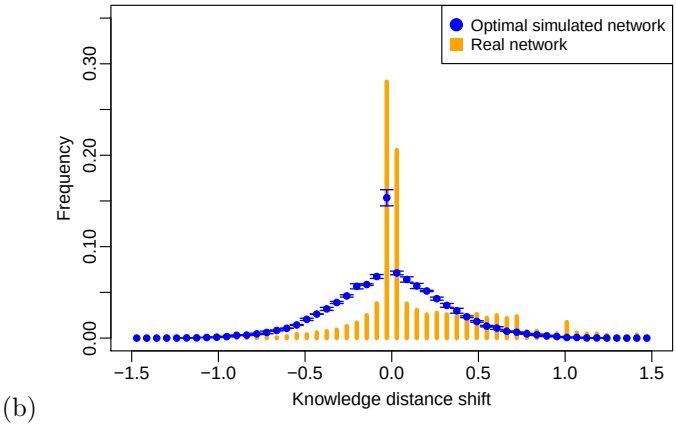


(b)

Figure 5.7: Empirical knowledge distance between every pair of partnered firms, computed 1, 3, 5 and 10 years after the date of the alliance formation. In (a) we have calculated the distance using the 8 dimensional knowledge space defined by the IPC scheme and in (b) used the 35 dimensional knowledge space defined by the ISI-OST-INPI classification scheme.



(a)



(b)

Figure 5.8: (a) Empirical shift of knowledge distance between every pair of partnered firms, computed 1, 3, 5 and 10 years after the date of the alliance formation. (b) Empirical and simulated distance shifts between all allied firms for $\tau = 700\text{days}$ and $\mu = 10^{-7}\text{days}^{-1}$. In both plots, we report results obtained considering the 35 dimensional space defined by the ISI-OST-INPI classification scheme.

Changes of knowledge distances In the second step, we calculate the *changes of the knowledge distances* between the empirical *pre-alliance* distance distribution shown in Fig. 5.2 and the empirical *post-alliance* distance

distribution shown in Fig. 5.7. Because the time of alliance termination is not known, we have to vary the duration again in time steps of 1 year. Our results are shown in Fig. 5.8 (a) for the ISI-OST-INPI classification scheme. The results for the IPC scheme are rather similar and therefore not shown here.

There are two remarkable observations in Fig. 5.8 (a). First, the distributions are clearly centered around zero, i.e. small changes of knowledge distances are very frequent. Larger changes of knowledge distances are rare, but not unlikely. This is in line with the broad distributions we find for both the pre- and the post-alliance knowledge distances. Second, the results for the changes in knowledge distances are robust against choosing a longer duration for alliances. We note that positive changes are more prominently seen for the ISI-OST-INPI classification scheme, whereas they look symmetric for the IPC scheme.

In order to see whether these findings are captured by our model of knowledge exchange, we have calculated the changes in distances also in the computer simulations (using optimal parameters). The result is shown in Fig. 5.8(b), where we compare the changes in the empirical knowledge distances (also shown on the left side) with the changes in the simulated knowledge distances. We see that the (rather) symmetric distribution peaked at zero can be reproduced by our model, even with the long tails. Some deviations occur close to zero, where the empirical distribution is more peaked, to decay faster than the simulated one. These deviations are in line with the deviations already discussed for Fig. 5.7, where small distances are slightly overrepresented in the simulated initial conditions.

More interesting is the fact that both the empirical and the simulated distributions of distance changes exhibit tails on *both* sides. I.e., some alliances cause the partners to significantly move *closer* in the knowledge space, whilst during other alliances the partners significantly move *farther away*. In our model of knowledge exchange, however, we have only assumed that alliance partners *approach* each other in knowledge space, which would lead to a left skew distribution of (mostly negative) changes. The explanation comes from the fact that firms, while forming new alliances, can be still engaged in existing alliances or establish new ones. The resulting change in the knowledge dis-

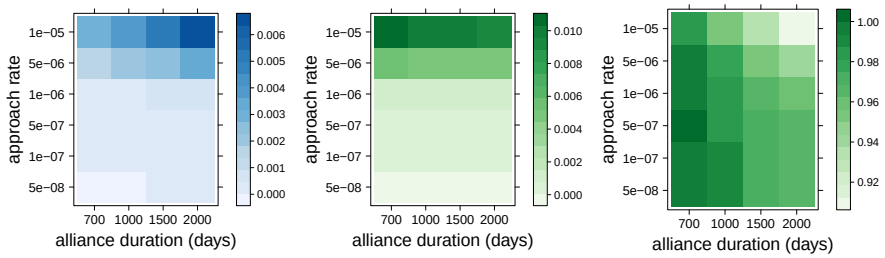
tance with respect to a given partner is thus the superposition of all influences a firm is subject to, at the time of alliance termination. In other words, there exists a nonlinear (and nontrivial) feedback of the network formation process on the knowledge exchange dynamics, which we further comment on in Sect. 5.6. At this point, we just emphasize that this influence is correctly captured in our agent-based model, as it also reflects the movement of agents farther away in knowledge space.

5.5 Estimating the performance of knowledge exchange

One of the most prominent reasons for R&D collaborations, seen from the perspective of the firm, is the exchange of knowledge, as already argued in Section 5.1. The formation of R&D alliances between individual firms results in a large-scale R&D network pictured in Fig. 5.1. This network represents one projection of the systemic, or "macroscopic", perspective. The complementary projection of the systemic perspective is given by the knowledge space made up by the patent portfolios of individual firms. Only the dimensions of that space are defined by the (two different) patent classification schemes. Firms collectively shape, and explore, this knowledge space by forming alliances and exchanging knowledge with their partners.

The collective exploration of the knowledge space is beneficial for the whole system [54]. Therefore, we now want to evaluate the performance of this collective exploration, by analyzing different indicators. We do not intend to directly match these indicators to any possible empirical counterpart. Rather, we address the question of to what extent the empirical R&D network corresponds to a simulated network that is optimal with respect to such indicators.

As the possibly simplest performance indicator for our simulated networks, we investigate the *total distance* that all agents have traveled in knowledge space [226]. For an individual agent, the length $L_i(t)$ of the path traveled in the



(a) (b) (c)

Figure 5.9: The heat map for the average total distance, $\langle L \rangle$, traveled by the agents is reported in (a). In (b) we report the heat map for network collaboration efficiency, $\mathcal{C}$, and in (c) the heat map for its normalized and rescaled version version, $\hat{\mathcal{C}}_n$. For all the three plots, we report results obtained using the 35 dimensional space defined by the ISI-OST-INPI classification scheme.

knowledge space is defined by the sum of all distances that the agent traveled in every time step of the simulation until time t:

$$L_i(T_{\max}) = \int_{t=0}^{T_{\max}} |\dot{\mathbf{x}}_i(t)| \, \mathrm{d}t \tag{5.6}$$

For our convenience $T_{\max}$ is the duration of the entire computer simulation. It should be noted that the measure $|\dot{\mathbf{x}}_i(t)| \, \mathrm{d}t$ is a positive scalar and expresses the actual *distance* traveled by the agent i, differently from its *net displacement* $\dot{\mathbf{x}}_i(t) \, \mathrm{d}t$, which is a vectorial quantity.

The measure $L_i(t)$ is then averaged over all the agents in the network to obtain the averaged total distance in knowledge space, $\langle L(t) \rangle = N^{-1} \sum_i^N L_i(t)$. This is shown in Fig. 5.9(a) as a heat map of the bi-dimensional $(\tau, \mu)-$parameter space. We argue that a higher value of $\langle L \rangle$, i.e. a better exploration of the knowledge space, corresponds to a higher systemic performance, because, as already discussed in Section 5.1.1, firms are proven to innovate more when they come in contact with more technological opportunities.

At the same time, using $\langle L \rangle$ as performance indicator does not give us detailed insights because, as Fig. 5.9(a) shows, higher approach rates μ always lead to

larger distances traveled in knowledge space, for any alliance duration τ. This motivates us to propose a more refined performance indicator, $\mathcal{C}$, that also takes into account the number of active collaborations, $k_i^{\mathrm{act}}(t)$, that cause firms to move in knowledge space at a given time t. I.e. in our model $k_i^{\mathrm{act}}(t)$ is the *degree* of agent i at time t. We remind that not all collaborations are active at a given time; some are terminated and become inactive after a characteristic time τ. As firms engaged in alliances incur in costs, we consider that $\mathcal{C}$ should decrease with increasing number of active collaborations:

$$\mathcal{C} = \int_{t=0}^{T_{\max}} \frac{\sum_{i=1}^{N} |\dot{\mathbf{x}}_i(t)|}{\sum_{i=1}^{N} k_i^{\mathrm{act}}(t)} \,\mathrm{d}t \;=\; \frac{1}{2} \int_{t=0}^{T_{\max}} \frac{\sum_{i=1}^{N} |\dot{\mathbf{x}}_i(t)|}{M^{\mathrm{act}}(t)} \,\mathrm{d}t \tag{5.7}$$

$\mathcal{C}$ is called *collaboration efficiency* because it considers how much output, i.e. movement in knowledge space, the system achieves for a given input, covering e.g. the costs to maintain collaboration links. The ratio of the two sums in Eq. (5.7) gives the total distance traveled *per active collaboration* during a given time step dt. This ratio is then integrated over the duration $T_{\max}$ of the simulation, to obtain the overall collaboration performance $\mathcal{C}$ of the network. The sum of all agents' degrees, $\sum_i k_i^{\mathrm{act}}(t) = 2 \cdot M^{\mathrm{act}}(t)$, gives us twice the total number of active links, $M^{\mathrm{act}}(t)$, in the network at time t because every link connects two agents. By plugging this into Eq. (5.7), we obtain the second expression for the collaboration efficiency. It means that, given equal total knowledge distances $\sum_i^N L_i(t)$, an R&D network with less alliances would explore the knowledge space more efficiently.

We use Eq. (5.7) to compute the collaboration efficiency $\mathcal{C}$ for every network generated during the simulations. $\mathcal{C}$ is then averaged over 100 simulations for every combination of parameters. The results are shown in the heat map of Fig. 5.9(b) for simulations using the 35 dimensional knowledge space, where we plot the collaboration efficiency $\mathcal{C}$ against the two parameters characterizing the knowledge exchange, exchange rate μ and alliance duration τ. Comparing this to Fig. 5.9(a), we find again that μ has a strong impact, i.e. the larger the knowledge exchange rate, the better the performance. However, the influence of τ has reversed. Now, performance increases with shorter alliance durations, which is understandable because we take the costs of alliances into account.

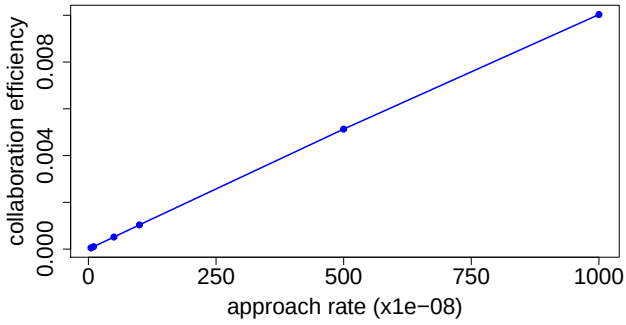


Figure 5.10: Collaboration efficiency $\mathcal{C}$ dependent on the knowledge exchange rate μ for a fixed alliance duration of $\tau = 700$ days. The knowledge of the agents was embedded in the 35 dimensional space defined by the ISI-OST-INPI classification scheme.

The larger τ, the more alliances exist concurrently and have to be maintained. This causes the overall performance to drop.

To further investigate the strong impact of μ, we plot in Fig. 5.10 for a fixed alliance duration τ =700 days how the collaboration efficiency $\mathcal{C}$ changes with the knowledge exchange rate. We find that there is a linear relation between these two quantities (similar for other values of τ, not shown). This is agreement with the definition of $\mathcal{C}$, Eq. (5.7), where the leading term of the numerator is linear in μ. Non-linear terms of the order $O(\mu^2)$ play a less important role since μ is small. Hence, for a better comparison of the collaboration efficiency across different values of τ, we rescale $\mathcal{C}$ as $\hat{\mathcal{C}} = \mathcal{C}/\mu$. In addition, to obtain a dimensionless quantity that varies between 0 and 1, we normalize $\hat{\mathcal{C}}$ by its maximum value obtained for a given set of parameters μ, τ, i.e.

$$\hat{\mathcal{C}}_n = \frac{\hat{\mathcal{C}}}{\max_{\mu,\tau} \hat{\mathcal{C}}} = \frac{\mathcal{C}/\mu}{\max_{\mu,\tau}(\mathcal{C}/\mu)} \tag{5.8}$$

In Fig.5.9(c), we show $\hat{\mathcal{C}}_n$ for all combinations of the knowledge exchange parameters. We confirm, even after normalization, the tendency that the performance increases with smaller values of τ, i.e., for the range of parameters

considered the best value of τ is 700 days. But for the knowledge exchange rate, we obtain a more detailed and heterogeneous dependency. Given τ =700 days, the optimal value of μ is now at 5×10^{-7} days^{-1}.

In conclusion, we find that the highest efficiency in knowledge exchange is obtained for medium exchange rates and short alliance durations. These results are found by means of computer simulations of our model. In order to transfer such insights to firms in real R&D networks, some restrictions apply.

It is understandable that a shorter collaboration is more beneficial because it implies, as already mentioned, that in a given time interval a smaller number of concurrent alliances exist. A reduced number of collaborations, on the other hand, allows a firm to move efficiently along one or a few directions in the knowledge space.

In order to keep the performance of exploring the knowledge space high, firms have to compensate shorter alliance durations by larger knowledge exchange rates. While this is feasible in our model, it may not hold under practical circumstances because firms have limits of how much new knowledge they can absorb at a given time. So, there are upper limits for the knowledge exchange rate μ.

On the other hand, it is obvious that there is a lower bound for an optimal alliance duration τ. Firms have to get to know each other, and have to establish procedures of collaborations which takes time. Hence, organizational and management arguments suggest that τ cannot simply approach zero, also because the knowledge exchange rate cannot simply be increased to arbitrary large values.

Such arguments apply when choosing *realistic ranges* of parameters τ and μ in our model. Thus, via the choice of parameters our model takes these limitations into account. In addition, it is useful to understand the *impact* of these model parameters on the systemic performance in knowledge exploration. As we have shown, there is a nonlinear, and non-trivial, relation between knowledge exchange rate μ and alliance duration τ. With an increasing alliance duration, more links become active at the same time, thus forcing firms to cope with the effect of multiple partnerships. This results in a reduced motion, i.e. a reduced collective exploration, in the knowledge space. In other words,

the density of the collaboration network increases with τ and, after a certain threshold, the addition of a new link has a negative marginal effect on the overall exploration of the knowledge space.

5.6 Conclusions

This chapter aims at a *quantitative* understanding of knowledge exchange in R&D networks. "Quantitative" means, (i) we propose a model that reflects the two tightly connected processes of forming R&D alliances and knowledge exchange, (ii) we analyze large-scale data sets capturing R&D alliances and knowledge bases of firms to calibrate the model parameters, (iii) we perform extensive computer simulations to analyze the performance of knowledge exchange in R&D network. Instead of repeating our findings, in this section we highlight a few points for further discussion.

Partner selection and network formation We have proposed an agent-based model that consists of two interlinked phases: (1) the formation of the R&D network, which is called the *exploration phase* because agents explore the social capital of potential partners, and (2) the exchange of knowledge on the formed network, which is called the *exploitation phase* because agents utilize the collaboration with partners to move in knowledge space.

The calibration of our model against real data was performed through a two-step procedure. By means of an alliance data set, we have estimated a set of link probabilities that allow us to reproduce the topology of the R&D collaboration network. Subsequently, through a second data set on firm patents, we have estimated parameters for the knowledge exchange between firms and the duration of R&D alliances.

For the formation of the R&D network, we found that firms exhibit a strong tendency to connect to network incumbents. Precisely, 65% of the collaborations initiated by incumbents, as well as a surprising 90% of the collaborations initiated by newcomers, are addressed to another incumbent. In this regard, the validation of our model brings additional support to the theory of the

importance of existing network structures in the formation of new R&D collaborations [see 165, 181].

Dynamics of knowledge exchange Because the model part related to the network formation was already investigated by Tomasello *et al.* [224], in this chapter we mainly focus on modeling knowledge exchange as a motion of agents in a predefined knowledge space. The knowledge base of agents is estimated by the patent portfolio of firms. Therefore, the dimensionality of the knowledge space is given by the patent classifications for which we use two different schemes, (a) IPC and (b) ISI-OST-INPI. With respect to our model, their difference is mainly in the *number* of dimensions, (a) 8 and (b) 35. Thus, we can also address the question how an expansion of the number of dimensions of the knowledge space affects the results of our model.

Firms are characterized by a position in this knowledge space, which changes over time as they obtain new patents. As at the focus of this chapter there are R&D *collaborations*, the model does not assume that firms can change their position by independent R&D activities. But we have indirectly covered this by the fact that, in our model, each time a new alliance starts agents get assigned a new position if they are not already involved in existing alliances. Differently from the model introduced by Tomasello *et al.* [226], where the motion of every agent was driven by only one partner at every time step, in the present model the agents are subject to a motion resulting from interactions with multiple partners. As we have already discussed in Sect. 5.3.2, our dynamics assumes that knowledge exchange causes agents to *approach* each other in knowledge space, not just in one dimension but in all dimensions. This takes into account the effect of knowledge spillovers that go beyond the exchange of very specific knowledge.

Analyzing *empirically* the impact of R&D collaborations on firms' knowledge positions, we found that small changes in knowledge distances are dominating the dynamics in knowledge space (see Fig. 5.8). I.e., real firms do not significantly change their knowledge positions as a consequence of their collaborations. This supports our conclusion that *most* alliances exert only a *weak* influence on the knowledge positions of firms. However, we also find

that *some* (non-negligible) alliances are able to cause a *strong* movement in knowledge space.

Interplay between network formation and knowledge exchange It is an interesting observation that the empirical distribution of distance changes is rather *symmetric* with respect to zero; although we note that positive changes are more prominently seen in the ISI-OST-INPI classification scheme (see Fig. 5.8). This means that, in the period elapsed during a specific R&D alliance, firms not only approach each other in knowledge space (negative distance changes) but also move farther away (positive distance changes).

This finding can be also reproduced by our agent-based model, which is remarkable because there we assume only that agents approach each other. However, the model of knowledge exchange considers the *combined impact* of all interactions on the knowledge position of an agent. Our model can reproduce both negative and positive distance changes because they result not only from the knowledge dynamics, but also from the *network dynamics*. This means that, while being engaged in one alliance, agents start to form new alliances with other partners which can drive them away from their current partners. Hence, it is the complex interplay between network formation and knowledge exchange that at the end can explain the collective exploration of the knowledge space.

Pre- and post-alliance distance distributions For the calibration of our knowledge exchange dynamics, special attention was given to the knowledge distances between firms at two points in time, at the moment of alliance formation (which is known) and at the moment of alliance termination (which is not known). Hence, the alliance duration τ is considered as one free parameter of our model.

We emphasize that in our model proximity in knowledge space is *not* a precondition for agents to form alliances. Consequently, distances can be quite *large*, which is in line with the empirical fact that the distribution of *pre-alliance distances* is clearly left-skewed (see Fig. 5.2). On the other hand, we have also shown that the most *frequent* pre-alliance distance between firms are shorter

than the one expected at random (see Fig. 5.4). The most probable value (i.e. the maximum of the distribution) is clearly different from zero and could be interpreted as an optimal distance in knowledge space for firms to engage in an alliance.

In our model, we have taken the distribution of pre-alliance distances as an input, i.e. we have sampled the knowledge positions of agents that are *not* engaged in an alliance at that time from this distribution. Agents that *are* in an alliance at that time, however, keep the knowledge position simulated by the model. The combined procedure of sampling knowledge positions has two advantages: first, we retain information about the similarity of collaborating firms in the knowledge space. For example, if firms from the same industrial sector were more likely to have an alliance, this would be captured in the pre-alliance distance distribution (e.g. smaller alliance distances are more probable) and considered in our model. Second, by using the empirical knowledge vectors, we also keep information about the technological areas in which firms usually file patents. Thus, we partially account for the size of firm portfolios of patents.

Regarding the distribution of *post-alliance distances*, we have shown that it is not really different from the distribution of pre-alliance distances, which reflects the fact that most changes in knowledge positions are rather small. This finding holds for both patent classification schemes, i.e. it is robust against the number of dimensions of the knowledge space. It is also robust against the assumed alliance duration (see Fig. 5.7).

So, if firms do not move much in knowledge space, why is their position important? Firms rather use the available information about knowledge positions of their partners to establish new collaborations. Therefore, a firm's position in knowledge space is more a *determinant* than a consequence of its R&D alliances.

In our model, we have used the distribution of post-alliance distances to compare the outcome of our simulations with their empirical counterpart. Using optimized parameters for the simulated network formation, we vary the parameters for knowledge exchange to find the best match between the empirical and the simulation post-alliance distance distribution (see Fig. 5.6). As the re-

sult, we obtain the values $\mu = 1 \times 10^{-7}, .., 5 \times 10^{-7}$ for the knowledge exchange rate and $\tau = 700$ for the alliance duration. μ has a relatively low value, which is in line with the fact that most firms do not move much in knowledge space, while τ indicates a characteristic duration of around two years (700 days). The latter finding is consistent with our previous theoretical assumptions and a number of previous studies [see 99, 163]. We note that these optimal parameters for knowledge exchange are obtained from a procedure that compares *simulation* and *empirics*. In the following, we discuss that we have derived the same optimal parameters from a pure simulation approach, using assumptions about performance.

Performance of knowledge exchange In this chapter, we are not only interested in the *dynamics* of knowledge exchange in R&D networks, but also in the *performance*. The latter we define as a systemic property, i.e. we do not discuss the performance of individual firms, but the collective performance of the whole R&D network in efficiently exploring the knowledge space.

The dynamics assumed for knowledge exchange would suggest that higher knowledge exchange rates μ and longer alliance durations τ are always better for exploration. This, however, implies that firms cope with many concurrent alliances at the same time and have an infinite capacity of absorbing new knowledge. A more realistic scenario has to take into account that alliances are also costly, i.e. establishing and maintaining concurrent alliances is constrained by capacities. To capture these influences, we have proposed the (normalized) *collaboration efficiency* $\hat{\mathcal{C}}_n$, Eqs. (5.7), (5.8), as a new performance measure. Analyzing how $\hat{\mathcal{C}}_n$ depends on the parameters for knowledge exchange μ and τ, we find that the collaboration efficiency is maximized for values $\mu = 5 \times 10^{-7}$ and $\tau = 700$ (see Fig. 5.9c), which match the above given optimized parameters from Fig. 5.6. Because this result was found by comparing only simulations, we regard this as an independent way to confirm the parameters found by comparing the empirical and the simulated distribution of post-alliance distances. This means that, using our approach, it is possible to obtain a configuration that is both *realistic* and *optimized* with respect to the collaboration performance.

When discussing these findings, we already pointed out that in real-world applications the parameters μ and τ are rather determined by the firm's abilities to quickly establish a collaboration and to absorb new knowledge fast. Hence, organizational and managerial constraints apply, which should be considered for choosing values for these parameters.

Nevertheless, with our model we are able to point toward policies aimed at system optimization. Effective policies to obtain an improved collaboration network would incentivize short R&D alliances and higher knowledge exchange rates. Practically, it would be impossible to directly enforce shorter alliance durations or faster learning rates. But measures could include, for instance, rewards for co-patenting activities if these are carried out as early as possible after the establishment of an R&D alliance. The goal would be to stimulate companies to explore other knowledge positions with new partners while limiting the duration of a single alliance and to avoid having too many active collaborations at the same time.

In conclusion, we argue that our model can successfully reproduce both network-related and knowledge-related features of a real inter-organizational R&D network. At the same time, our data-driven approach provides a unique method to estimate the systemic performance of R&D collaborations. We note that our model is extendable to other collaboration systems, beyond the domain of R&D networks, provided that the agents can be unequivocally positioned in a knowledge space. Our approach thus contributes to a comprehensive understanding of the effects of knowledge exchange in dynamically evolving collaboration networks.

Chapter 6

Scientists' knowledge distance and knowledge exchange

Summary

We analyze a large publication data set to study to which extent knowledge is exchanged between scientists using collaborations. We start by defining a knowledge space using a real-world classification scheme for publications. In this space, we assign knowledge positions to scientists depending on their publications. Then, we analyze how scientists knowledge positions change over time, i.e., how scientists move in the knowledge space depending on newly co-authored publications. Also, we analyze scientists' productivity with respect to the knowledge of their set of collaborators. Finally, we reconstruct the aggregated collaboration network between scientists and identify a relation between scientists' distances on the network and in the knowledge space.[1]

[1]This chapter contains unpublished work. Recently, we have discovered that our definition of the knowledge space was used by [103] to model the evolution of scientists' interest publishing in the APS data. Additionally, Jia *et al.* [103] have chosen the cosine distance to measure knowledge distances in this space, just like we do. At the same time, the methods and focus of this chapter have significant differences compared to [103]. In particular, from the method perspective, we do not normalize the size scientists' knowledge positions, and hence, we do not re-scale the knowledge space. In Sect. 6.3, we discuss why we do not have to normalize in our study. Moreover, in [103], the authors have a different focus compared to us. They mainly concentrate on changes in scientists' positions in the knowledge space between the beginning and the end of their careers. We provide an extensive analysis of the

6.1 Introduction

Globalization is leading to an increase of collaborations in many areas of human activity, ranging from economics to science [85, 142]. In the previous chapter, we have analyzed collaborations in economics, and now we turn our attention to science. In this domain, the increase of collaborations is quantitatively captured by an increasing number of publications co-authored by teams of scientists [70, 125, 256]. By working together, scientists combine their different knowledge and strengths in order to produce new results. These results are then reported in co-authored scientific publications that encode new knowledge jointly produced. This is mainly possible thanks to a co-ordination process that involves communication and knowledge sharing between scientists. Hence, the analysis of co-authorship activities provides not only an understating of how knowledge is jointly produced but also of how it is exchanged.

Many works have analyzed scientists' co-authorship activities and have mainly focused on the formation of individual collaborations and on how a network structure emerges from these. Examples include the pioneering network analysis of Newman [143, 145, 149] that have characterized scientist collaboration networks, and the models for team assembly of scientists proposed by Guimera *et al.* [81] and Tomasello *et al.* [227]. However, less attention has been devoted to the role of knowledge in the formation of collaborations and on how these collaborations then affect the knowledge of the collaborating scientists. A notable exception is [219] where the authors have proposed a model for the evolution of scientific disciplines depending on scientists' co-authorship activities. Regardless of all these advances, the questions of how collaborations influence the knowledge of individual scientists and how collaborations are dependent on scientists' knowledge are still open.

In this chapter, we provide an extensive empirical analysis of scientists' collaborations and of the feedback between these collaborations and scientists' knowledge. To do this, we start by embedding scientists in a *knowledge space* depending on their co-authored publications listed in the APS corpus (see

feedback between scientists' positions and collaborations instead. All the work presented in this Chapter was developed independently of [103].

Chap. 2). The knowledge space is defined using the PACS classification scheme that is used to classify these publications. In this space, we introduce and justify a *knowledge distance* capturing how similar or dissimilar scientists are depending on their knowledge. Then, we use the introduced knowledge space to study scientists' collaborations at three different levels.

First, we provide an analysis at the micro-level where the units of analysis are pairs of scientists. By this, we extend the analysis of Chap. 5 to the scientific domain. Also, we introduce two new measures: one capturing scientists' effort in reducing the knowledge distance with their collaborators and one quantifying knowledge exchange between scientists that have co-authored a publication.

Second, we explore collaborations and knowledge at the meso-level. By referring at groups of scientists co-authoring publications as teams, we study the composition of teams with respect to the knowledge of their members. Moreover, we investigate how the productivity of scientists is dependent on the knowledge of their set of collaborators.

Third, we move our attention on a macro-level and analyze the entire scientists' collaboration network. This network is reconstructed by aggregating all the co-authorship activities listed in our data. We characterize the reconstructed network by computing standard network measures and verify that our network is similar to other collaboration networks studied in previous works. Then, we analyze how scientists' distances on the network are related to their distances in the knowledge space.

The remaining of the chapter is divided as follows. In Sect. 6.2, we shortly describe the data used. In Sect. 6.3, we define a knowledge space using the PACS classification scheme, knowledge positions of scientists, and knowledge distances. Then, we analyze scientists' collaborations at a micro-level in Sect. 6.4, at a meso-level in Sect. 6.5, and at a macro-level in Sect. 6.6. Finally, we conclude this chapter with Sect. 6.7 where we discuss our findings.

6.2 Data

We determine the knowledge space and knowledge positions of scientists using the APS data set. This contains more than 460 000 publications coming from 1893 to 2009 (116 years). For each publication, we have various information, including its list of authors, PACS codes, DOI, and publishing date. From the DOI, we obtain disambiguated scientists' information using the procedure described in Chap. 2. Note that we have the PACS codes only for papers published after 1975, i.e., from the year PACS was introduced. Hence, we discard older publications.

We also discard publications with more than 40 co-authors. Recall that we consider co-authoring activities as collaboration events that allow the exchange of knowledge. In other words, we assume that all the scientists co-authoring a publication have been interacting with each other. With increasing group size, the probability that these interactions have really occurred decreases. As an example, consider large physics experiments, like LIGO, whose results are published in papers co-authored by hundreds of scientists. However, probably only specific sub-groups of scientists worked together. For this reason, we remove publications co-authored by too large groups. Note that it is essential to remove these publications from the analysis. Each collaboration of size m provides $m(m-1)/2$ pairwise interactions at scientist level. So 9 900 papers with only two co-authors (for which it is reasonable to assume a close collaboration among the co-authors) give the same number of pairwise interactions as two papers with 100 authors. This means that any statistics computed using pairwise interactions would be extremely dependent on publications co-authored by large groups. For this reason, we remove these publications.

Additionally, we also discard scientists that have either authored only on publication or only single-author publications. For the former type of scientists, we cannot analyze their evolution of time, while for the latter, we have no information about their collaborators.

After the above filtering procedure, the final sample used for this analysis contains 109 845 scientists authoring 299 052 publications. We represent scientists and collaboration using a network perspective. By this, we reconstruct the

collaboration network among scientists publishing in APS journals. The network is undirected and unweighted, and it has 109 845 nodes and 1 232 539 distinct links.

6.3 Preliminaries

6.3.1 The knowledge space

To represent the knowledge of a scientist, we use the publications that he/she has co-authored. We do this by embedding scientists in a knowledge space defined using a real classification scheme of publications: the Physics and Astronomy Classification Scheme (PACS). The PACS contains codes that are assigned to publications listed in the APS data. The PACS codes are made of pairs of two digits number and succeed by two characters: upper or lower case letters or "+" or "-" symbol. For example, a PACS code is *13.30.Er*. The first digit of the first pair of digits ("1") represents the main category: "1" stands for `The Physics of Elementary Particles and Fields`. The first and second digit together still define broad sub-categories: "13" is `General theory of fields and particles`. The second pair of digits and the combination of characters represent a much more narrow and specialized classification: "11.30.Er" is `Charge conjugation, parity, time reversal, and other discrete symmetries`. [2]

To define the dimensions of the knowledge space, we have three options: to use only the first digits of the PACS codes, use the first pair of digits or use the full codes. Using only the first digits would imply that we have only ten different types of topics. This would give to much of a coarse-grained view of the different types of knowledge contained in the publications. While using the full codes would give an extremely sparse representation. We argue that by keeping the first two digits, we can capture enough differences while retaining enough statistics. By using the first two digits, we obtain a total

[2]The full description of the different available PACS codes are in `https://publishing.aip.org/publishing/pacs/pacs-2010-regular-edition`.

First digit	PACS name	# of Second digits
00	General	7
10	The Physics of Elementary Particles and Fields	4
20	Nuclear Physics	8
30	Atomic and Molecular Physics	7
40	Electromagnetism, Optics, Acoustics, Heat Transfer, Classical Mechanics, and Fluid Dynamics	7
50	Physics of Gases, Plasmas, and Electric Discharges	2
60	Condensed Matter: Structural, Mechanical and Thermal Properties	8
70	Condensed Matter: Electronic Structure, Electrical, Magnetic, and Optical Properties	9
80	Interdisciplinary Physics and Related Areas of Science and Technology	7
90	Geophysics, Astronomy, and Astrophysics	8
TOTAL		67

Table 6.1: First digit of the PACS codes with their name and their number of distinct second digit codes.

of **67 different** topics that will represent the 67 different dimensions of the knowledge space (see Table 6.1).

Note that the PACS is not constant over time, but we do not consider this in our study. We argue that to include the time evolution is rather a complication than of any help. PACS has been continuously evolving between 1985 to 2010 [158]. To consider the evolution of PACS codes would cause complications in the classification of the publications and the determination of the knowledge space, while bringing unclear benefits. Only for some particular cases, changes in the classifications scheme are relevant, such as when a PACS code is removed. For example, PACS code 22 was removed in the last version of the classification. Since we have only two papers with such code, we have discarded them from our analysis and this PACS code 22 is not considered as a dimension in the knowledge space. For most codes, the classification scheme did not undergo profound changes, especially at the second digit level. Therefore we discard the time evolution of the PACS. For an interesting study on the evolution of the PACS using a network perspective, see [158].

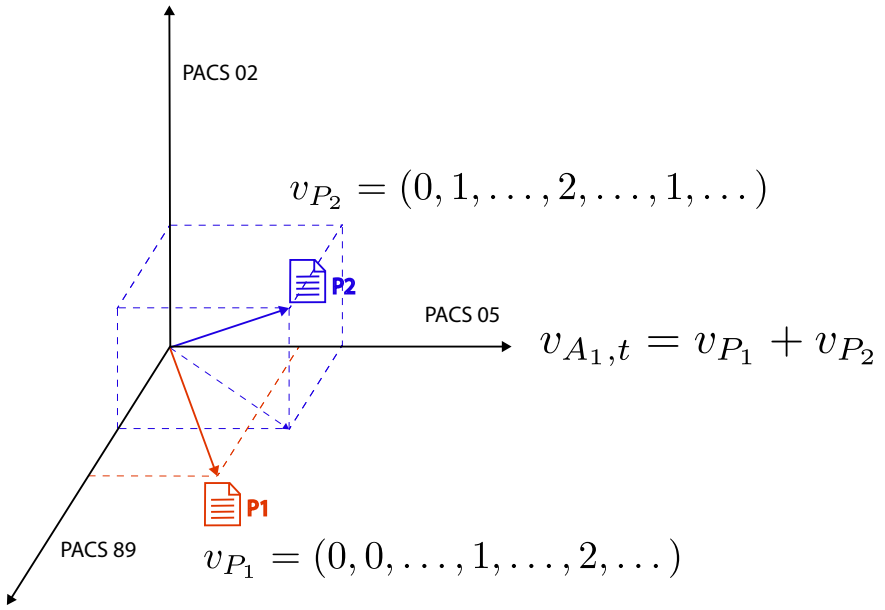


Figure 6.1: Knowledge vectors of two papers, v_{P_1} and v_{P_2}, authored by the same author A_1 before time t. The knowledge vector of the author $v_{A_1,t}$ is equal to the sum of the knowledge vectors of the two papers.

6.3.2 Knowledge positions of scientists

Each scientist is placed in the knowledge space depending on its publications. Recall that this comes from the convention that publications are artifacts that encode explicit knowledge (see Chap. 1). Hence, we start by assigning to each publication a knowledge vector depending on its PACS codes. For example, if a publication α has three PACS codes **02**.10.*Ab*, **03**.65.*Aa* and **03**.65.*Ca*, we assign a knowledge vector $v_\alpha = (0, 1, 2, ...)$: each component of v_α is equal to the number of PACS codes assigned to the publication α and related to that component (see Fig. 6.1). Note that the knowledge vectors of publications are vectors of integers.

After defining knowledge vectors of publications, we define *scientists knowledge positions* as the sum of the knowledge vectors of the publications that

he/she has authored alone or in collaboration. In other words, a scientist i has published a set $S_i(t)$ of papers until time t, i.e. $S_i(t) = \{p_\alpha, p_\beta, \dots\}$, and for each paper, p_α, we have assigned a knowledge vector v_α. Then, the knowledge position of scientist i at time t is

$$v_{i,t} = \sum_{p_\gamma \in S_i(t)} v_\gamma \tag{6.1}$$

Scientists trajectories. By looking at scientists' positions at different times, we reconstruct their *trajectories* in the knowledge space. The first feature of such trajectories is that they occur on a multidimensional lattice as changes of positions depend on publications' knowledge vectors that contain only integers number. Also, trajectories always move away from the origin of the knowledge space. In such a space, a scientist can quickly move away from the origin if he/she authors publications always with similar PACS code. While, if he/she continuously authors publications with different PACS codes, he/she explores more dimensions of the space and stays closer to the origin. We call the former an exploration "in-depth" of the knowledge space, while the latter an exploration "in-breath". The difference between "in-depth" and "in-breath" exploration can be captured by computing the euclidean distances between scientists' positions and the origin after a given number of publications. In other words, by computing the euclidean length of their trajectories.

In Fig. 6.2, we report the empirical distribution of lengths after publishing 2, 10, 20 and 50 papers. We find that the distributions are quite broad and have an increasing median of 3.46, 14.73, 28.27, and 67.01. This corresponds to having half of the scientists authoring publications in at least 3, 6, 9 and 14 PACS (respectively after 2, 10, 20 and 50 publications). From this, we find that scientists tend to focus on specific topics at the begging of their careers, but then over time, they increasingly diversify their production. In order to quantitatively address this point, we compare the lengths of the empirical trajectories with the one generated with a null model. The null model is based on a random walk process in the knowledge space. Each scientist is replaced with a random walker, and the random walker obtains the publication list of the scientist. At each time step, the random walker takes a publication

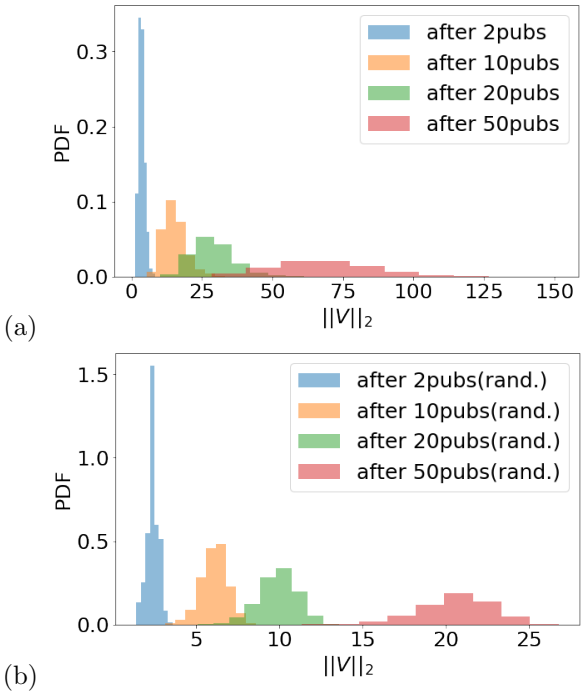


Figure 6.2: Distribution of euclidean norms of the empirical knowledge positions (a) and of the random ones (b) after 2 (blue), 10 (orange), 20 (green), 50 (red) publications.

from her list and performs a number of jumps equal to the number of PACS codes present on the publication. By this, we simulate trajectories for which we can compute the lengths. We find that the random walkers explore many more dimensions and hence, move much more slowly away from the origin compared to the real scientists. The medians for the simulated distributions of trajectories are 2.45, 6.16, 10.0 and 20.76, corresponding to having half of the random walkers exploring at least 5, 23, 38 and 59 dimensions, respectively after 2, 10, 20 and 50 publications. Additionally, most of the random walkers have trajectories with similar lengths, while the distributions of trajectories' lengths for the real scientists are much broader.

To summarize, we find that scientist explore the knowledge space in an exceptionally non-random way with respect to two aspects: 1) each scientist has some preferred directions, i.e., topics; 2) the movement behaviors of scientists in the knowledge space are much more heterogeneous compared to what would be expected from a random walk process. To further study scientists' trajectories in the knowledge space and to capture the above aspects, we could consider using reinforced random walks [47]. In this type of random walk, the dynamic depends on the previous history, and they have been successfully applied to model directed motions of humans and animals [93, 209]. Thus, we could try to reproduce scientists' trajectories by correctly tuning the importance of previously explored dimensions of the knowledge space. Even though such a modeling perspective is interesting, it deviates from the focus of this chapter, which is studying knowledge exchange occurring during co-authorship activities of scientific publications. For this reason, we leave as future research to verify how scientists trajectories could be modeled using reinforced random walks.

6.3.3 The knowledge distance

In order to measure if two scientists have similar or dissimilar knowledge, we introduce a knowledge distance in the knowledge space. We choose to use the *cosine distance*:

$$d_{cos}(v_1, v_2) = 1 - \frac{v_1 \cdot v_2}{|v_1||v_2|}\,. \tag{6.2}$$

where v_1 and v_2 are vectors identifying knowledge positions of two scientists. This measure varies between 0 (scientists have similar knowledge) and 1 (scientists have dissimilar knowledge). Precisely, $d_{cos} = 0$ when two scientists have parallel vectors. This happens if and only if both scientists published in the same PACS in the same proportion and hence, they have knowledge on the same topics in the same proportion. If two scientists have half of their publications containing the same PACS codes in the same proportion, then their cosine distance is $1 - \sqrt{2}/2 \approx 0.3$ and corresponds to an angle of 45 degree between their knowledge positions v_1 and v_2. For angles smaller than

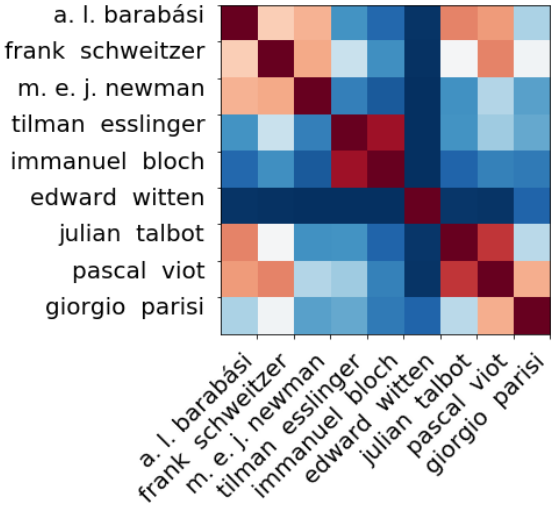


Figure 6.3: Distance matrix among nine established scientists publishing in APS. A red color indicates that two scientists are similar, while a blue color indicates that they are dissimilar.

45 degrees, scientists have similar knowledge as they share at least half of their knowledge. For bigger angles, scientists have dissimilar knowledge.

A case study. To provide a better intuition of what knowledge distance captures, we make a case study. We have selected nine established scientists publishing in APS journals in different sectors for whom we can judge how similar and dissimilar they are from our personal experience. The set of authors are A. L. Barabási, Frank Schweitzer, M. E. J. Newman, Tilman Esslinger, Immanuel Bloch, Edward Witten, Julian Talbot, Pascal Viot, Giorgio Parisi. In Fig 6.3, we report their knowledge distances in 2010 as a distance matrix. Each cell represents the similarity between two scientists, and the colors go from red (similar) to blue (dissimilar). In the top left corner, we have a similarity between A. L. Barabási, Frank Schweitzer, and M. E. J. Newman. These scientists research on complex systems and network theory, and they are cor-

rectly mapped to have similar knowledge ($d_{cos} \approx 0.3$). Two among these three scientists are measured to have similar knowledge to Julian Talbot and Pascal Viot that are working on statistical physics. Given the fact that complex systems and statistical physics are two overlapping topics, our measure correctly estimates the similarity between them. From our measure, we also find that Edward Witten is dissimilar from all the others, and this matches the fact that he is one of the top scientists working on string theory. To our knowledge, none of the others are working on this topic [3]. Last, but not least, we look at Tilman Esslinger and Immanuel Bloch that are two renowned scientists working on remarkably similar topics: cold atoms and optical lattices. Our measure captures their similarity as their knowledge distance is almost equal to zero ($d_{cos} \lesssim 0.1$). Note also that cold atoms and optical lattice physics are topics that do not overlap with the ones of the other scientists in our sample. Thus, Tilman Esslinger and Immanuel Bloch have dissimilar knowledge to the other scientists ($d_{cos} > 0.8$).

In Chap. 5, we have used the Euclidean distance d_E, and hence, one could argue to use this measure again, instead of the cosine distance. However, with d_E, we would be capturing two difference in scientists knowledge at the same time: 1) differences in the number of published papers and 2) differences in the type of knowledge. For example, consider two scientists, i and j, that always publish in one single PACS, but one has 1 paper, while the other 6 papers. If we assume that each paper only has one PACS, then knowledge distance between i and j is $d_E(v_i, v_j) = 5$ and it only comes from the different number of published papers. This knowledge distance could been obtained also by two scientists k and l that published 3 and 4 papers in different PACS, $d_E(v_k, v_l) = \sqrt{3^2 + 4^2} = 5$). This time, the knowledge distance is capturing that the authors are publishing on different topics. Since we only focus on detecting differences in the type of knowledge, i.e., the second scenario and not the first one, we have to fix this ambiguity.

[3]The only exception is Giorgio Parisi whose primary research focus is spin glasses (statistical Mechanics), but he also had some significant contribution in QCD and string theory. Our measure partially captures this as Giorgio Parisi is the less dissimilar scientists from Edward Witten.

Note that the ambiguity arises from the fact that now the knowledge positions of scientists are not normalized, while in Chap. 5 firms' knowledge positions were normalized. However, this time, we do not normalize the vectors of knowledge positions as the lengths of these vectors are an exciting feature of our data (see in Sect. 6.3.2). So, we have decided to use the cosine distance (see Eq. (6.2)). This is almost equivalent to use the Euclidean distance on normalized vectors. Precisely, the cosine distance between two vectors, v_1 and v_2, is proportional to the square of the euclidean distance between their respective *versors* (i.e., their normalized vectors, $\hat{v}_i = v_i/|v_i|$ with $i = 1, 2$):

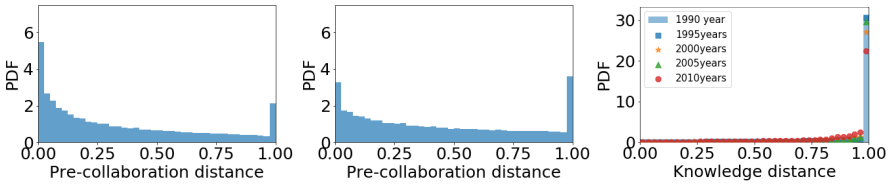
$$d_{cos}(v_1, v_2) = 1 - \hat{v}_1 \cdot \hat{v}_2 = \frac{1}{2}\left(\sum_i (\hat{v}_1)_i^2 + \sum_i (\hat{v}_2)_i^2\right) - \sum_i (\hat{v}_1)_i (\hat{v}_2)_i$$

$$= \frac{1}{2}\left(\sum_i (\hat{v}_1)_i^2 + (\hat{v}_2)_i^2 - 2(\hat{v}_1)_i (\hat{v}_2)_i\right) = \frac{1}{2} d_E^2(\hat{v}_1, \hat{v}_2)$$

where $\hat{v}_j = v_j/|v_j|$, i.e. a versor, $(\hat{v}_j)_i$ is th i-th component of the knowledge versors v_j and we have used that $1 = \sum_i (\hat{v}_l)_i^2$. Thus, when comparing authors using the cosine distance, we are actually using an equivalent measure to the one used in the previous chapter. [4]

After introducing the knowledge space, scientists' knowledge positions and how to calculate distances, we have the right tools to quantify the volutions of scientists' knowledge. In the next two sections, we analyze the interplay between scientists' knowledge positions and their collaboration patterns. The analysis is split into three parts. First, we analyze pairwise interactions, i.e., similarities and differences between pairs of collaborating scientists. Second, we focus on teams, meaning that we analyze how groups of scientists have similar or dissimilar knowledge when co-authoring the same publications. Third, we look at the entire collaboration network.

[4]There is only one small difference between these two measures in our case. We have normalized the knowledge vectors of firms using an L1 norm and not the Euclidean, L2. By using the L1 norm, it was easier to interpret the meaning of the component of the knowledge vector of a firm. They are the fraction of patents filed in a specific IPC code.



(a) (b) (c)

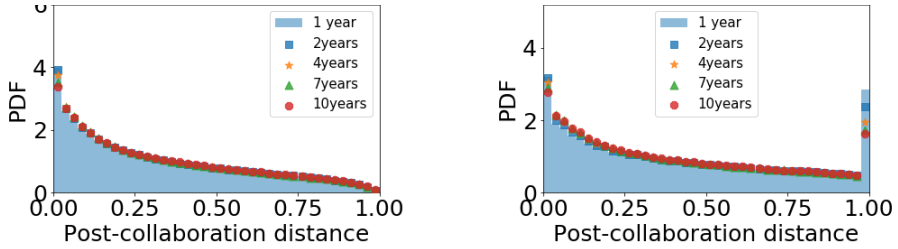
Figure 6.4: Distributions of pre-collaboration distances before every single collaboration (a), before collaborations among distinct scientist pairs (b) and among random pairs of scientists (c). For this last distribution, we look at distances in five different years 1990, 1995, 2000, 2005, 2010.

6.4 Knowledge distance and knowledge exchange: A micro-level analysis

We study the effect of collaborations on scientists' knowledge positions. Similar to the Chap. 5, we start to do this by investigating three different quantities: pre-collaboration distances, post-collaboration distances and knowledge shifts (see Sect. 6.4.1). Then, we study how collaborations influence scientists' knowledge exchange and their approach (or distancing) in the knowledge space (see Sect. 6.4.2).

6.4.1 Distribution of knowledge distances

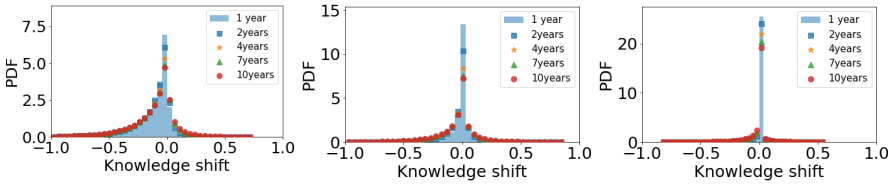
Pre-collaboration distance. We start by looking at the distribution of knowledge distances between pairs of authors before they decide to collaborate. This distribution tells us how similar or dissimilar co-authors are. In other words, what type of co-authors are usually preferred. In Fig. 6.4(a), we report such distribution. We find that there are two noticeable picks: one at little knowledge distances and one at high knowledge distance. This means that authors usually prefer authors who are similar or dissimilar. The pick at small knowledge distances can be understood from repeated collaborations, and indeed, if we consider only the first time two authors have collaborated then, this peak becomes quite smaller (Fig. 6.4(b)). While the second pick at $d_{cos} = 1$ shows that often authors prefer to collaborate with extremely dif-



(a) (b)

Figure 6.5: Distributions of post-collaboration distances among distinct scientist pairs. In particular, in (a) the knowledge vectors of the scientists are constructed using their full publication list. While in (b), we discard the knowledge coming from the paper that they have just co-authored. We look at post-collaboration distances after five different time windows of length equal to 1, 2, 4, 7, and 10 years.

ferent authors. Recall that a cosine distance equal to 1 means that the two authors have orthogonal knowledge vectors, i.e., they have never published in the same PACS. One could argue that this effect is dependent on missing information, i.e., that many authors have only authored one paper and we do not have enough information to assign them precise knowledge positions. However, when correcting for this, meaning that we consider only pre-collaboration distances among scientists with at least two publications, we still have the pick (not shown). Moreover, note that the distribution of pre-collaboration distances obtained cannot be expected at random. By computing the knowledge distances between random pairs of scientists, we obtain the expected distribution of knowledge distances (under a random null model). In Fig. 6.4(c), we report this distribution and find that most scientists listed in our data have orthogonal knowledge among each other. This clearly shows that knowledge positions of scientists and their decision with whom to collaborate are coupled events. We find that the mean and median of the three distribution are 0.34 and 0.24 when considering every pre-collaboration distance (Fig. 6.4(a)), 0.43 and 0.37 when considering only the first co-authored publications (Fig. 6.6(b)), ~ 0.90 and 1.0 when considering random pairs of scientists (Fig. 6.4(c).



(a) (b) (c)

Figure 6.6: Distributions of knowledge shifts among distinct scientist pairs. In particular, in (a) the knowledge vectors of the scientists are constructed using their full publication list. While in (b), we discard the knowledge coming from the paper that they have just co-authored. We look at post-collaboration distances after five different time windows of length equal to 1, 2, 4, 7 and 10 years. In (c), we plot the distribution of knowledge shifts of random pairs of scientists.

Post-collaboration distances. The post-collaboration distances are computed between scientists co-authoring a publication. In Fig. 6.5(a), we report the post-collaboration distances computed by comparing knowledge positions of scientists after 1, 2, 4, 7, and 10 years from their first collaborations. We find that from the mean and median of the post-collaboration distances are 0.32 and 0.25 that are quite smaller compared to mean (0.43) and median (0.37) of the pre-collaboration distances reported in Fig. 6.6(b). Part of this drop is given from the fact by construction post-collaboration distances of length 1 are not allowed. Indeed, after co-authoring a paper, two authors cannot have distances equal to one, i.e., orthogonal vectors in the knowledge space. To correct for this, in Fig. 6.5(b), we report the post-collaboration distances computed by comparing scientists' knowledge positions from which we have removed their first common publication. We see that this distribution has a pick at 1. Note also that this distribution is more similar to the empirical pre-collaboration distributions (see Fig. 6.4(b)), but its mean and median (~ 0.39 and ~ 0.33) are still smaller compared to the pre-collaboration distributions. Last, to check that distributions of post-collaboration (Fig. 6.5(a) and (b)) could not be expected at random, we can compare them to Fig. 6.4(c) From a visual comparison, it is clear that these are really different distributions.

Knowledge shifts. We now consider the change in the knowledge distance between two scientists after co-authoring a publication, i.e., after collaborating. We name this quantity knowledge shift as in Chap. 5. If we observe a negative knowledge shift between two scientists, we learn that their knowledge distance is decreasing and hence, their knowledge is becoming more similar. While if their knowledge shift is positive, then their knowledge is becoming more dissimilar. In Fig. 6.6(a), we report the knowledge shift distributions computed after 1, 2, 4, 7, and 10 years. In addition, we report the same distributions computed by excluding the first paper that two scientists have co-authored together in Fig. 6.6(b). Also, we compute the knowledge shifts coming from random pairs of authors and report this third distribution in Fig. 6.6(c). We find that the knowledge shifts tend to be centered in zero for all distributions, but the right-hand side of the distribution is heavier for the no-random pairs, i.e., Fig. 6.6(a) and (b). This means that most scientists move closer to each other in the knowledge space after collaborating, but there are also rare cases where they move away from each other.

6.4.2 Knowledge effort and knowledge exchange

In the previous section, we have analyzed the relative changes in knowledge positions between pairs of scientists. We further extend our analysis by introducing two new measures. The first measure quantifies how much a scientist is moving towards or away from his/her collaborators in the knowledge space. We name this first quantity *knowledge effort* as it captures scientists' effort to decrease or increase his/her distance from his/her collaborators. The second measure that we introduce quantifies *knowledge exchange*.

Knowledge effort. We have observed that most pairs of scientists move closer to each other after their first collaborations (see Fig. 6.6). Then, we ask: Who makes an effort to decrease the knowledge distance? Is it a joint effort or one scientist move more than the other? To answer these questions, we break down the knowledge shifts in four contributions to identify the different efforts that scientists make. Recall that the pre-collaboration distance between two scientists with knowledge vectors, v_1 and v_2, is the cosine dis-

tance: $d_{cos}(v_1, v_2) = 1 - \frac{v_1 \cdot v_2}{|v_1||v_2|}$. After a certain time Δt, the scientists might have published new papers and have new knowledge vectors, $v_1' = v_1 + \Delta_1$ and $v_2' = v_2 + \Delta_2$. Their post-collaboration distance is:

$$d_{cos}(v_1', v_2') = 1 - \frac{v_1' \cdot v_2'}{|v_1'||v_2'|} = 1 - \frac{v_1 \cdot v_2 + v_1 \cdot \Delta_2 + v_2 \cdot \Delta_1 + \Delta_1 \cdot \Delta_2}{|v_1'||v_2'|} \quad (6.3)$$

By taking the difference $d_{cos}(v_1', v_2') - d_{cos}(v_1, v_2)$, we obtain the knowledge shift between the two scientists:

$$\Delta_{1,2} = \frac{v_1 \cdot v_2}{|v_1||v_2|} - \frac{v_1 \cdot v_2 + v_1 \cdot \Delta_2 + v_2 \cdot \Delta_1 + \Delta_1 \cdot \Delta_2}{|v_1'||v_2'|} = c_1 + c_2 + c_3 + c_4 \,. \quad (6.4)$$

where

$c_1 = -\frac{v_1 \cdot \Delta_2}{|v_1'||v_2'|} \leq 0$, directed effort done by the scientist-2 to become more similar to scientist-1

$c_2 = -\frac{\Delta_1 \cdot v_2}{|v_1'||v_2'|} \leq 0$, directed effort done by the scientist-1 to become more similar to scientist-2

$c_3 = -\frac{\Delta_1 \cdot \Delta_2}{|v_1'||v_2'|} \leq 0$, common effort done by both scientists to become more similar to each other

$c_4 = \frac{v_1 \cdot v_2}{|v_1||v_2|} - \frac{v_1 \cdot v_2}{|v_1'||v_2'|} > 0$ as $|v_i'| \geq |v_i| > 0 \, \forall i$

Note that the first three terms are always smaller than or equal to zero and they decrease the knowledge shift, while the fourth term can only increase it. We name c_1 and c_2 the *directed efforts* of scientists as they are quantify whether scientist-2 moves towards scientist-1 or viceversa. Then, we name c_3 the *common effort*. Two scientists decrease their knowledge distance ($\Delta < 0$) if and only if their directed and common efforts balance and overtake c_4 ($|c_1 + c_1 + c_1| > c_4$) .

In Fig. 6.7, we report the cumulative distribution of directed efforts, c_1 and c_2, and of common efforts, c_3, after 1, 4, and 10 years. We find that for scientists' pairs experiencing negative knowledge shifts, *both* the directed and common efforts are more negative. We quantify this by doing a Kolmogorov-Smirnov-test between the distribution of efforts for negative and positive knowledge shifts.

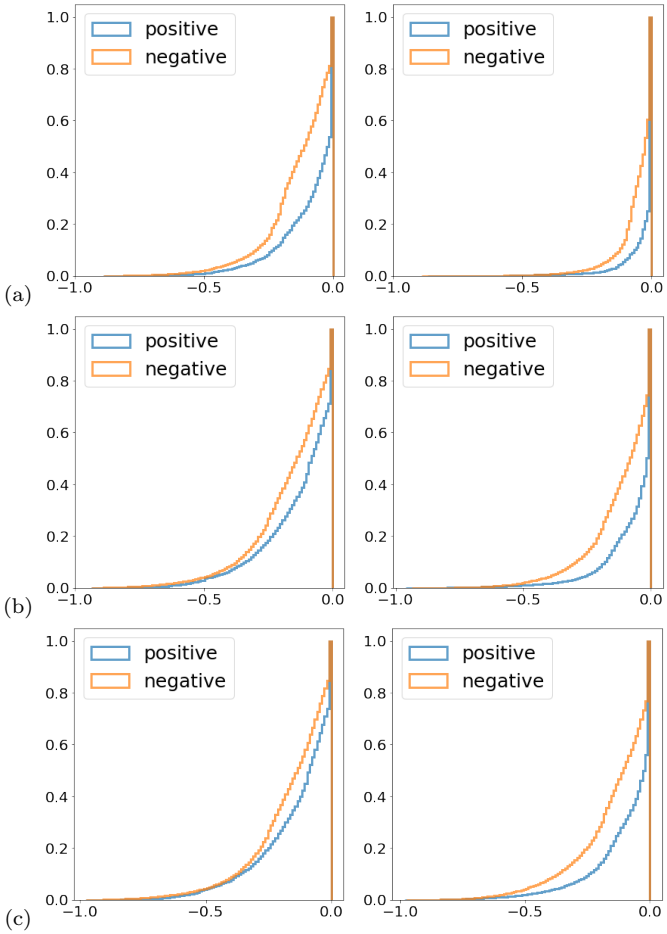


Figure 6.7: Cumulative distribution function of the directed knowledge efforts (left) and the common knowledge efforts (right) after 1 (a), 4 (b), and 10 years(c). In each plot, we present the distributions for scientist pairs with positive (blue) and negative (orange) knowledge shifts. In other words, in blue, we have pairs of scientists that become more different from each other after their first collaboration, while in orange, we have pairs of scientists becoming more similar.

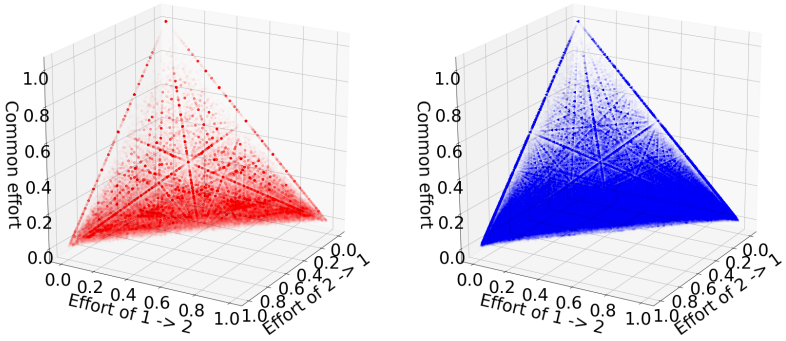


Figure 6.8: Scatter plots of the relative contributions of the efforts to the knowledge shift after 1 year. We use red circles when $\Delta > 0$ (scatter plot on the left) and blue triangles when $\Delta \leq 0$ (scatter plot on the right).

We find that for all the cases we cannot reject the null hypothesis that the distributions are different with a confidence bound of 1%. This finding means that the distributions of directed and common efforts for negative knowledge shifts are different from the distributions of efforts for positive knowledge shifts.

In addition, note that the difference between the distributions in Fig. 6.7(left) become more similar over time, i.e., the orange and blue curves get closer to each other. This implies that the directed efforts for positive and negative knowledge shifts become more similar over time. While common efforts stay more negative only for the scientists having a negative knowledge shift. Thus, just after collaborating, two scientists decrease their knowledge distance as at least one of them makes an effort to decrease the gap. Then, over the years, knowledge distances are decreased as both scientists publish papers in the same PACS (i.e., on the same topics).

To further compare the contributions of directed and common efforts (c_1, c_2 and c_3) to knowledge shifts, (Δ), we use a 3D scatter plot in Fig. 6.8. In this figure, each point has co-ordinates $(-c_1/\delta, -c_2/\delta, -c_3/\delta)$ where $\delta = |\Delta - c_4|$. We use red circles when $\Delta > 0$, i.e. for scientists moving away from each other (Fig. 6.8 (left) and Fig. 6.9 (left)), and blue triangle when $\Delta < 0$, i.e. for

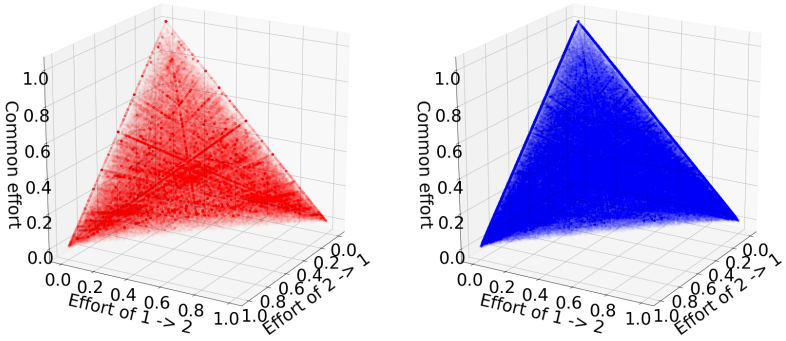


Figure 6.9: Scatter plots of the relative contributions of the efforts to the knowledge shift after 10year. We use red circles when $\Delta > 0$ (scatter plot on the left) and blue triangles when $\Delta \leq 0$ (scatter plot on the right).

scientists approaching each other (Fig. 6.8 (right) and Fig. 6.9 (right)). From these figures, we see that the top tip of the triangular surface is quite empty (Fig. 6.8) and over time it gets more filled (Fig. 6.9) for both negative and positive knowledge shift. By this, we find that just after collaborating, the more significant contribution to decrease knowledge distances comes from the directed efforts of one of the two scientists. Only after many years, common efforts become more important in decreasing knowledge distances.

Knowledge exchange. To quantify the knowledge exchange between two scientists, we calculate how much changes in the knowledge position of a scientist is aligned to the knowledge of the other scientist. When collaborating, scientists have access to knowledge that they might have previously not had. If a scientist *re-applies* this new knowledge to produce new publications, then we argue that some knowledge has been exchanged.

Using the same notation introduced to define knowledge effort, we define v_1 and v_2 to be the knowledge positions of two scientists at time t, and Δ_1 and

Δ_2 to be their change is knowledge positions after a time Δt. Now, we split the change in position of scientist-1 into two terms:

$$\Delta_1 = \Delta_{1,//} + \Delta_{1,\perp} \tag{6.5}$$

where $\Delta_{1,//} = (\Delta_1 \cdot \frac{v_1}{|v_1|})\frac{v_1}{|v_1|}$ and $\Delta_{1,\perp} = \Delta_1 - \Delta_{1,//}$. We name $\Delta_{//}$ the *aligned change* (in position) as it is the change in position of scientist aligned to his/her previous knowledge While $\Delta_\perp$ is the change in position of scientist perpendicular to his/her previous knowledge and we name $\Delta_{1,\perp}$ the *perpendicular change* (in position).

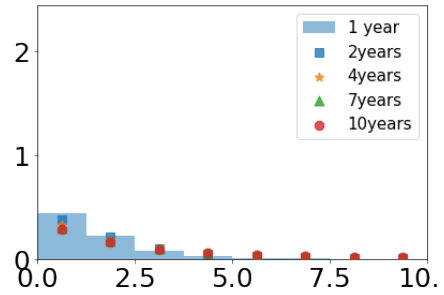
Assuming that scientist-1 and scientist-2 have collaborated at time t, we define the knowledge exchange from scientist-2 to scientist-1 (after a time Δt) the perpendicular change in position of scientist-1, $\Delta_{1,\perp}$, aligned to the knowledge of scientist-2 v_2 (at collaboration time):

$$\text{know.-ex.}_{2\to1} = \Delta_{1,\perp} \cdot \frac{v_2}{|v_2|} \tag{6.6}$$

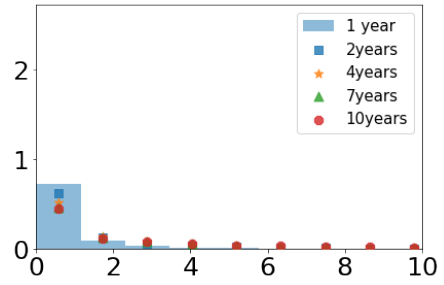
Note that if scientist-1 publishes one new paper with new knowledge (e.g., $\Delta_\perp = (1, 0, \ldots, 0)$) that is similar to the knowledge of one of its collaborator (e.g., $v_2 = (1, 0, \ldots, 0)$), then know.-ex.$_{2\to1} = \Delta_{1,\perp} \cdot \frac{v_2}{|v_2|} = 1$. Hence, when know.-ex.$_{2\to1}$ is number bigger than one, we have a good overlap between the new knowledge produced by scientist-1 and the knowledge of scientist-2. This indicates a high knowledge exchange. If $\Delta_\perp = (1, 0, \ldots, 0)$ and $v_2 = (1, 1, \ldots, 0)$, then we have that only one dimension of the new knowledge of scientist 1 is aligned to the knowledge position scientist 2. In this case, know.-ex.$_{2\to1} = \sqrt{2} \approx 0.707$ and hence, we argue that for know.-ex.$_{2\to1} \in [\sqrt{2}, 1]$ we have a relative high knowledge exchange. While for know.-ex.$_{2\to1} \in [0, \sqrt{2})$, we have low knowledge exchange from scientist-2 to scientist-1.

Likewise, the knowledge exchange from scientist-1 to scientist-2 (after a time Δt) is the perpendicular change in position of scientist-2, $\Delta_{2,\perp}$, aligned to the knowledge of scientist-1 v_1 (at collaboration time):

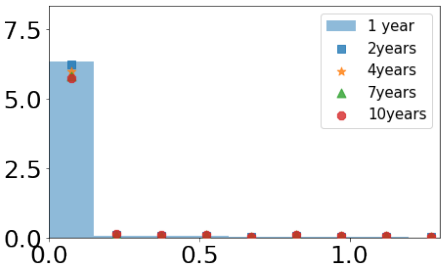
$$\text{know.-ex.}_{1\to2} = \Delta_{2,\perp} \cdot \frac{v_1}{|v_1|} \tag{6.7}$$



(a)



(b)



(c)

Figure 6.10: Distributions of knowledge exchange among distinct scientist pairs. In particular, in (a) the knowledge vectors of the authors are constructed using their full publication list. While in (b), we discard the knowledge coming from the paper that they have just co-authored. We look at collaboration distances after five different time windows of length equal 1, 2, 4, 7, and 10 years. In (c) we plot the distribution of knowledge exchange among 5000 random pairs of scientists.

In Fig. 6.10(a), we report the distributions of knowledge exchanges know.-ex.$_{i\to j}$. We find that the values of such quantities are quite large, and the mean and median are: 0.73 and 0.32 after 1 year, 0.97 and 0.40 after 2 years, 1.36 and 0.51 after 4 years, 1.79 and 0.59 after 7 years, 2.08 and 0.63 after 10 years. To see whether this strong knowledge exchange depends on the first paper that two scientists co-author together, we recompute knowledge exchanges removing this type of paper. In Fig. 6.10(b), we report the distributions of recomputed knowledge exchanges know.-ex.$_{i\to j}$. We find that the values of such quantity are quite smaller and now rarely different from zero as for only about 40% of scientists pairs we have know.-ex.$_{i\to j} \neq 0$: Now the mean is 0.29, 0.54, 0.94, 1.37 and 2.66 respectively after 1, 2, 4, 7, 10 years. This indicates that there is a small quantifiable knowledge exchange among 40% of scientist pairs if we correct for the first paper co-authored together.

As robustness check of the above result, we verify whether the observed values of knowledge exchange can be expected at random. To do this, we randomly pair scientists and compute the distribution of knowledge exchange between them (Fig. 6.10(c)). We obtain a mean of 0.04, 0.08, 0.15, 0.25, and 0.36 respectively after 1, 2, 4, 7, and 10 years, and the medians are always equal to zero. Additionally, we find that know.-ex.$_{i\to j} \neq 0$ for less than 20% of the random pairs. Hence, the values of knowledge exchange expected at random are much smaller compared to the empirical ones and much more often equal to zero. Therefore, the observed distributions of knowledge exchange (Fig. 6.10(a) and (b))could not been obtained at random and are statistically significant.

6.5 Knowledge in teams and productivity: A meso-level analysis

After analyzing knowledge distances and exchange between *pairs* of scientists, we now look at the role of knowledge inside groups of scientists co-authoring publications. We refer to such groups as *teams* because we assume that co-authoring scientists collaborate and share the common goal of publishing the paper.

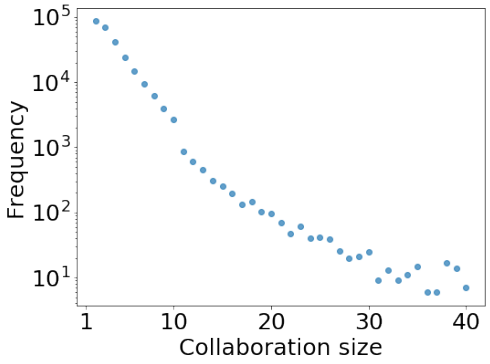


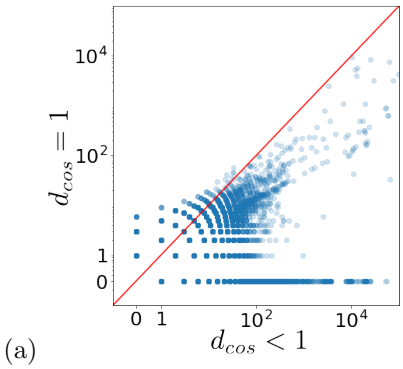
Figure 6.11: Frequencies of team size. The maximum size is 40 as we discard publications that are co-authored by more than 40 scientists.

By analyzing teams, we move our analysis of collaborations from the micro-level to the meso-level. At this level, we have to keep in mind two essential aspects. The first one is that collaborations do not have a homogenous structure. With this, we mean that each team is composed of a different number of scientists that have diverse knowledge. The second aspect to keep in mind is that every scientist can work in different teams over the years. Hence, scientists' knowledge and productivity might be dependent on the teams in which they work. We analyze these two aspects in Sect. 6.5.1 and 6.5.2.

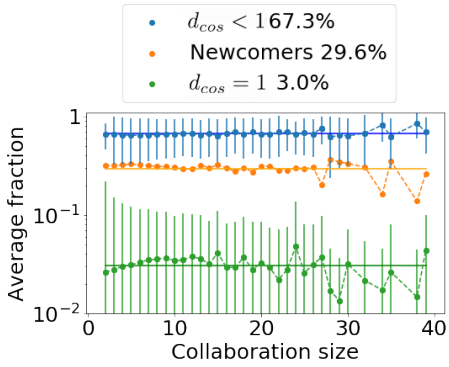
6.5.1 Team composition

In Fig. 6.11, we report the frequencies of team sizes. We find that even though most publications have less than ten co-authors, there are still thousands of publications co-authored by larger teams. Recall that the larger collaboration size is 40 as we have discarded larger collaborations (see Sect. 6.2). Also, note that the scale of the y-axis is in log scale.

In Sect. 6.4.1, we have found that most scientists collaborate with other scientists with similar and overlapping knowledge. This was captured by having the majority of pre-collaboration distances smaller than 1 (see Fig. 6.4(a)). At the same time, a considerable large fraction of pairwise collaborations is



(a)



(b)

Figure 6.12: (a) Number of links established between scientists with overlapping (x-axis) and orthogonal (y-axis) knowledge. (b) Fraction of links per collaboration established between scientists with overlapping (blue) and orthogonal (green) knowledge in function of the collaboration size. In orange, we report the fraction of links per collaboration where for one the two scientists were a newcomer. We also report error bars, except for the fraction of newcomer to keep the graph more readable.

established between scientists without overlapping knowledge, i.e. by having pre-collaboration distances equal to 1. Inside a team, we find that both these two types of pre-collaboration distances are present. We report this in 6.12(a) using a scatter plot. Each point is a team, and its x and y-coordinates are

the number of pairwise interactions in the team between scientists with and without overlapping knowledge.

From Fig. 6.12(a), we also find another interesting phenomenon. With increasing team size, fewer and fewer teams are composed of scientists without overlapping knowledge. We argue that this arises from two facts. First, when a team is bigger, there is a smaller probability (at random) not to have at least two scientists sharing some knowledge. Second, in larger teams coordination efforts increase as more scientists have to communicate their work with each other. Assuming that communication between scientists is more efficient among scientists sharing similar knowledge, we can understand why it is highly un-probable that large teams are composed only by scientists with very dissimilar knowledge ($d_{cos} = 1$).

To further characterize teams, we plot the average fraction of links (per team) established between scientists with overlapping ($d_{cos} < 1$) and not overlapping knowledge ($d_{cos} = 1$) in Fig. 6.12(b). We find that independently of the team size, about 67% of the links are established between authors with pre-collaboration distances smaller than 1. Only 3% of the links are between scientists with pre-collaboration distances equal to 1. The remaining 30% are links involving "newcomers", that are scientists for which we have not yet assigned a knowledge position. Note that in Chap. 5 we have named newcomers firms that have not yet participated in alliances. Similarly, here we call newcomers those scientists that are co-authoring a publication for their first time.

We have observed that there are different types of pre-collaboration distances inside a team independently of its size. Now we investigate how this affects the average pre-collaboration distances inside a team. In Fig. 6.13, we report the average pre-collaboration distance in function of the team size. We find that the average knowledge distances among co-authors are almost constant until the team size reaches 10. After 10, it starts decreasing. We argue that this phenomenon matches the idea that to have different types of knowledge inside a team is beneficial, however only until a certain point. Each scientist has a limited amount of energy that he/she can use to bridge the knowledge differences with its collaborators. More collaborators he/she has in a team,

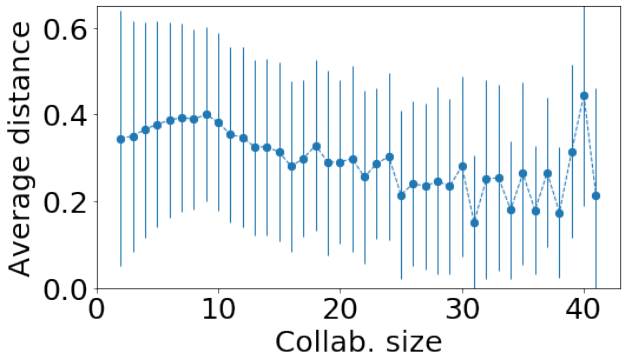


Figure 6.13: Average pre-collaboration distance in function of the collaboration size. The error bars represent the standard deviations among the average pre-collaboration distances at fixed collaboration size.

less energy on average he/she can use for each of them. However, given the big error bars, a detailed statistical analysis is necessary to verify the above statement. We leave such an analysis for future work.

6.5.2 Productivity and knowledge breath

The main scope of scientific co-authorship is the production of knowledge artifacts, i.e. publications. As we noted in Sect. 6.3.2, scientists can either publish papers in the same PACS(s) or publish papers in many different ones. These two types of behaviors can be interpreted as two different strategies for being productive. To study these strategies, we estimate scientists' productivity using two quantities. The first is the total number of publications co-authored by a scientist. While the second is the *knowledge breath* of a scientist.

We define the *knowledge breath* of a scientist, $k_{i,t}^{breath}$, as the change in position of a scientist after Δt years. We quantify this by computing $k_{i,t}^{breath}(\Delta t) = d_{cos}(v_{i,t}, v_{i+\Delta t})$. This captures scientists' exploration of the knowledge space by addressing new topics. Indeed $k_{i,t}^{breath}(\Delta t) = 0$ when scientist-i co-authors publications always in the same PACS (i.e., in the same dimension) during the

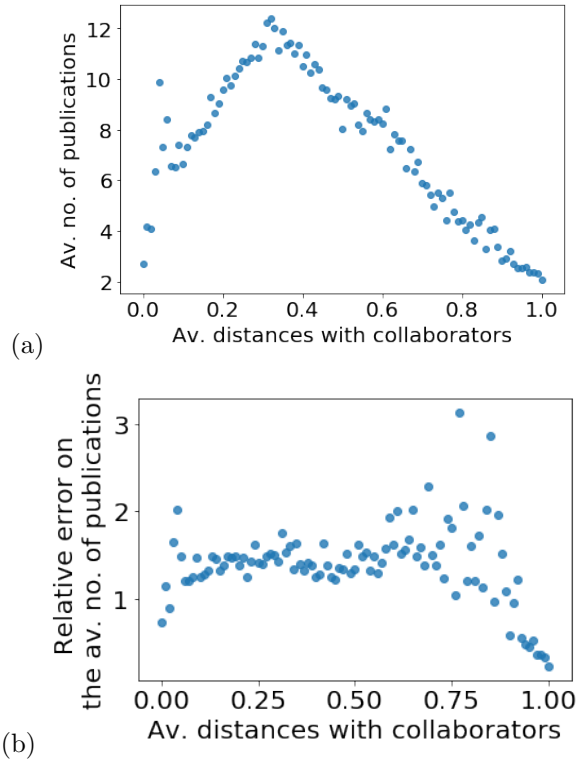


Figure 6.14: (a) Average number of publications and (b) its relative error in function of the pre-collaboration distances.

time window Δt. While $k_{i,t}^{breath}(\Delta t) \neq 0$ if he/she co-authors publications in different PACS.

In Fig. 6.14, we report the average number of publications per scientist as a function of the average pre-collaboration distances with their collaborators. We find that the more productive scientists are the ones collaborating with other scientists at an average pre-collaboration distance of about 0.35. In Fig. 6.15, we report the average knowledge breath of scientists as a function of the average pre-knowledge distances of their collaborators. We find that the knowledge breath has an inverted "U" shape after 1, 2, 4, 7, and 10 years from the collaboration. In particular, the maximum average knowledge breath is

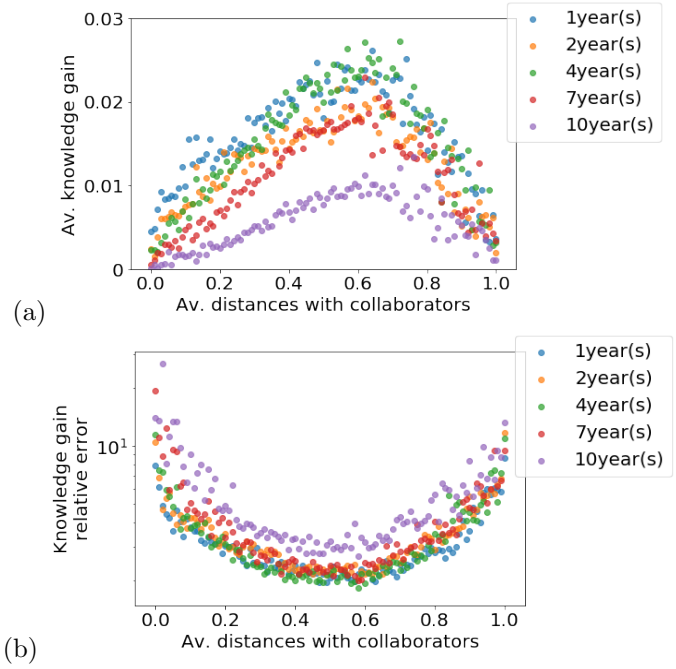
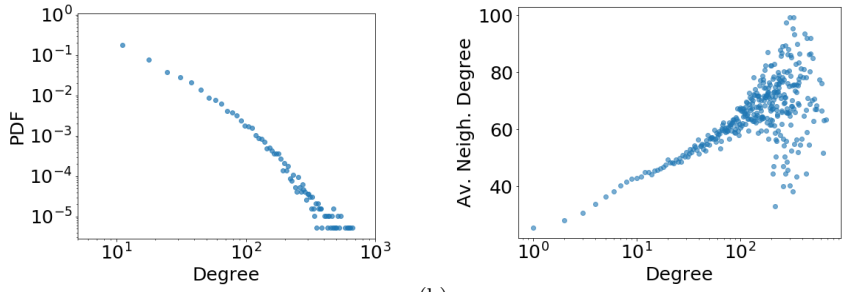


Figure 6.15: (a) *Knowledge breath* and (b) its relative error in function of the pre-collaboration distances.

observed for scientists collaborating with other scientists at an average knowledge distance of about 0.60.

To summarize, we find that scientists maximize their productivity and knowledge breath when choosing collaborators at two different pre-collaboration distances. Scientists involved in more collaborations have their collaborators at an average pre-collaboration distance of 0.35. While scientists with a higher knowledge breath, they collaborate with scientists at an average pre-collaboration distance of 0.60.



(a) (b)

Figure 6.16: Distribution of degree (a) and neighbor connectivity (b).

6.6 The collaboration network and knowledge distances: A macro-level analysis

To complete our analysis of scientists' collaborations, we switch our attention to the entire *collaboration network*. As discussed in Sect. 6.2, this network is composed of 109 845 nodes, representing scientists, and 1 232 539 links, representing their collaborations. In order to understand the characteristics of this network, we start by calculating standard measures common in network analysis. This includes the *degree distribution* $P(d)$ where d is the number of collaborators of a given scientist. In Fig. 6.16(a), we report the degree distribution and find that it is broad with a median of six and an average degree of 12.85. This means that most of the scientists over the years collaborate with other six scientists. Note that Newman [147] also computed the average number of collaborators using co-authorship activities in different disciplines. He obtained slightly less than four for theoretical disciplines and above 15 for experimental disciplines. In our data, we have groups working on both theoretical and experimental disciplines and hence, it is reasonable to obtain an average value between the ones obtained by Newman [147].

In Fig 6.16(b), we report the *neighbor connectivity* that measures to what extent scientists are linked to other scientists with a similar degree. We find that scientists with a lower degree tend to prefer to co-author publications with scientists with a higher degree. With increasing scientists' degrees, this

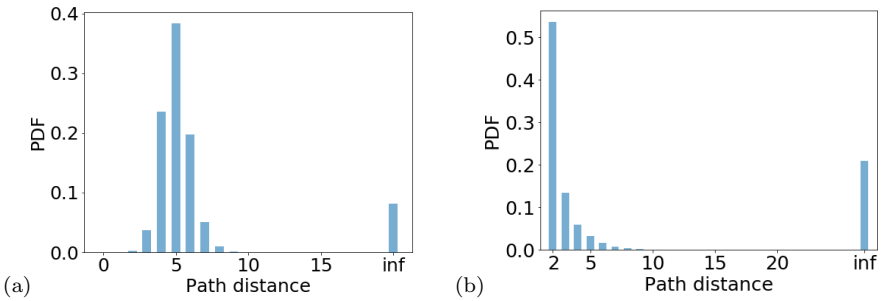


Figure 6.17: Distribution of shortest distances between 1mio random pairs of scientists (a) and between the observed pairs of scientists before they collaborate (b). For (b), we consider only the first time a pair of scientists collaborate, i.e., we do not consider repeated interactions. This is why there are no network distances equal to 1.

tendency becomes less strong. Overall, we find that the collaboration network has a positive assortativity coefficient, $r = 0.125$. Such value is in close with the ones reported for various collaboration networks re-constructed using co-authorship activities in [148].

In Fig. 6.17(a), we report the distribution of *shortest distances* between 1mio random pairs of scientists on the collaboration network. We find that shortest lengths are peaked around five and less than 10% of the pairs are disconnected. This means that five steps are necessary to move from one random scientist to another. This is in line with the results reported by [145]. Additionally, in Fig. 6.17(b), we plot the distribution of *shortest distances* between pairs of scientists before they collaborate. We find that most of the collaborations are established among scientists that have a small distance on the collaboration network. Only extremely rarely, collaborations are established between far-away authors. This result is also in line with our previous findings reported in [227].

In Fig. 6.18(a), we report the knowledge distances as a function of the shortest distances between every pairs of scientists on the time-aggregated network. The x-coordinate of each point represent the shortest distance between pairs of scientists before they collaborate. While the y-coordinate is the average

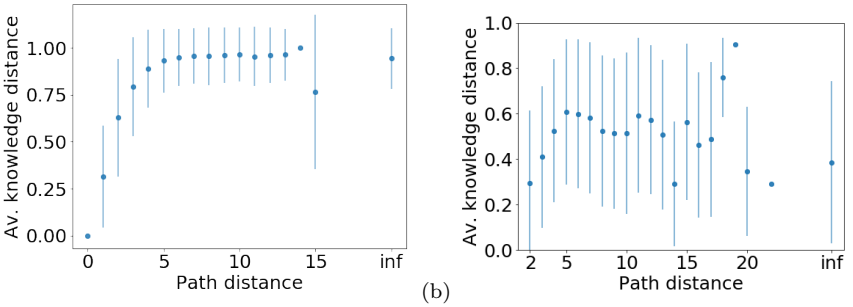


Figure 6.18: Pre-collaboration distances in the knowledge space as a function of the pre-collaboration distances on the collaboration network. (a) We consider all the collaboration listed in our data set. For (b), we consider only the first time a pair of scientists collaborate, i.e., we do not consider repeated interactions. This is why there are no network distances equal to 1.

pre-collaboration distance between pairs of scientists at a given network distances. We find that scientists' distances in the knowledge space increase with increasing network distance. Note that scientists five steps away from each other (the most frequent distance) are usually at a knowledge distance close to one. This confirms our previous result that most of the authors have a high knowledge distance (see Fig. 6.4(c)).

Also, Fig. 6.18(b), we plot the knowledge distances as a function of the shortest distances between pairs of scientists before they collaborate. We find that for scientists close in the collaboration network have similar knowledge ($\langle d_{cos} \rangle \approx 0.3$). The average knowledge distance increases to a maximal value of around 0.6 for scientists at network distance between five and seven. This shows that scientists not only prefer collaborators close in the collaboration network but among them, they prefer the ones that are closer in the knowledge space.

6.7 Conclusion

We have performed an extensive analysis of scientists' collaborations and how these influence scientists' knowledge. Our analysis was divided into three parts: a micro-level analysis where the units of analysis are scientists pairs, a meso-

level analysis where the units of analysis were teams, and a macro-level analysis where we considered the full collaboration networks. Instead of repeating the different results, we now conclude by making a connection to our results obtained in Chap. 5 and in other previous works.

6.7.1 Comparing firms and scientists

We now aim to provide a comparison between firms and scientists. On the one hand, one can argue that firms and scientists are different actors embedded in extremely different environments, namely the economic and academic domain. Hence, we should not compare them. On the other hand, we are interested in analyzing scientists and firms with respect to one specific aspect: the possible knowledge exchange triggered by collaborations. Indeed, to analyze this aspect in both the economic and academic domains, we have used remarkably similar data sets and methodologies.

We have used IPC and PACS, two classification schemes, respectively for patents and publications, to construct a knowledge space. To embed firms and scientists in their respective knowledge spaces, we have used patent and publication data, i.e., knowledge artifact data. We have introduced two distance measures: one to quantify differences in firms' knowledge positions, and a second one to quantify differences in scientists' knowledge positions. As discussed in the Sect. 6.3.3, these distances are almost equivalent. For both domains, we have computed the empirical distributions of pre-collaboration, post-collaboration distances, and knowledge shifts. Recall that we have named these distributions pre-alliance, post-alliance, and knowledge shift distributions in the economic domain (see Chap. 5). By all this, we have analyzed using the same methods the interplay between knowledge and collaborations in the economic and academic domains. Hence, our methodology allows us to have a direct and reasonable comparison across domains.

Pre-collaboration distances. This distribution in the academic domain is bimodal: it is picked at zero and 1. Its counterpart in the economic domain has only one pick at a particular distance. Thus, we argue that this reflects two different behaviors. Firms prefer to collaborate with other firms that have

a different knowledge base, but not too different otherwise the communication would become too difficult. While a scientist has to more extreme behaviors: either exploiting his/her own expertise by working with other experts from his/her fields, or exploring different fields by collaborating with experts with a different knowledge base. On the one hand, it is surprising to find such extremely different behaviors from the economic and the academic domain, especially after showing their collaboration patterns could be explained using the same modeling approach [227]. On the other hand, we have to recall that scientists and firms are indeed extremely different actors. In addition, they have different reasons not only to collaborate but also to publish/file knowledge artifacts. Hence, the detected difference in the pre-collaboration distances is a valuable source of information.

Post-collaboration distance. In the economic domain, the distribution of pre-alliance distances is extremely similar to the post-alliance one. In the academic domain, these two distributions are profoundly different. The distribution of post-collaboration distances cannot contain knowledge distances equal to 1, i.e., two collaborators cannot have orthogonal knowledge vectors. Recall that we observe a collaboration among two scientists when they co-author a paper and hence, both their future knowledge vector will share the knowledge of co-authored paper. Sharing this common knowledge makes scientists' knowledge distance smaller than 1. However, when correcting for this, i.e., we compute the knowledge vectors of two collaborators discarding their *first* joint publication, we find that the distributions of post-collaboration distances and pre-collaborations distances are actually similar. To summarize, for both domains, we find that the knowledge position of the authors is slowly changing.

Knowledge shift. For both the domains, we find that the distributions of knowledge shifts are centered in 0 and take negative and positive values. However, for the academic domain, the negative side of the distribution is much heavier. This observation is robust also when calculating post-collaboration distance among scientists without considering the first joint publication. We argue that this difference arises from the fact that for scientists publishing work

on a new topic is more accessible. In other words, for a scientist to acquire and produce knowledge on a topic in which he/she is not an expert is easier compared for firms. Additionally, in the economic literature co-patenting is also seen "as a second-best option that, if possible, should be avoided" [86]. Also for this reason, we can argue that the negative knowledge shifts are harder to detect in the economic domain compared to the academic one.

Note that we used similar types of data to calculate the pre- and post- collaboration distances in the two domains, but they have some differences. For the economic domain, we used two separated data sets to calculate knowledge positions and analyze collaboration patterns. While in the academic domain, we infer collaborations and knowledge using the same data. This implies that scientists after a collaboration cannot have a maximal distance anymore. We have corrected for this by studying the post-collaboration distances between scientists discarding the (first) co-authored publications.

To embed firms and scientists in a knowledge space, we used real classification schemes PACS and IPC. The main difference between PACS codes compared to the IPC is that scientists assign PACS codes to their work, while patent reviewers assign IPC coded to firms' patents. This means that the assignment of each code is decentralized and dependent on each scientist. At the same time, the APS journals are peer-reviewed and hence, all publications with their PACS codes are controlled.

As future research, we leave the validation of the knowledge effort and exchange measures, and their application in the economic domain. These measures quantified how scientists move close or away from each other and how much knowledge is transferred after a collaboration. We have shown that this measure gives results that cannot be expected at random and hence, provide a good starting point for quantifying the influence co-authorship activities of scientists' knowledge. However, they still need to be validated against other data sets or via scrutiny of experts. Additionally, their application in the economic domain is still open.

6.7.2 Knowledge distances as a proxy for communities

In Sect. 6.5.1, we have obtained that the teams are composed by scientists with both overlapping ($d_cos < 1$) and not overlapping knowledge ($d_cos = 1$). In particular, when representing a team with a fully connected clique, 67% of links are between scientists with overlapping knowledge, 3% of links are between scientists not overlapping knowledge, and the remaining 30% of links involve a newcomer (i.e., a scientist publishing his first paper). Note that it is quite striking to find such extreme differences in the preferences to be stable over team sizes, ranging from 2 to 40.

The obtained fractions of links match previous results about scientists' preferences in choosing collaborators [227]. In this other work, we have used the agent-based model for network formation described in Sect. 5.3.1 to reproduce scientists' collaboration networks. We have found that scientists choose collaborators belonging to their community with a high probability ($p_s^L \approx 0.62$) to a different community with a really low probability ($p_d^L \approx 0.05$), and to newcomers with a probability of $p_n^L \approx 0.33$ [5]. Note that in [227], communities are implicitly defined through shared practices and/or behaviors and are assigned by our ABM. Hence, they are created without any information about scientists' positions in the knowledge space. By this, we have an un-expected match between the empirical analysis presented in this section and our ABM.

We argue that the match found arises from the fact that knowledge distances are a good proxy for communities that are strong determinants of collaborations. By this, we mean that the community of a scientist defined in [227] contains scientists at a $d_{cos} < 1$. While scientists at $d_cos = 1$ belong to different communities. Hence, we can use our measure of knowledge distances to define communities that we can then use in our ABM to determine teams of possible collaborators. This idea is also supported by the obtained relations between knowledge and path distances on the collaboration network. In particular, we have found that scientists choose collaborators that are not only

[5] The value reported in the present has been obtained by taking the average of the values of p_s^L reported in Table 2 of [227]. When calculating the average we have discarded only the values coming from the co-authorship network for general relativity and gravitation (PACS 04). For this PACS, the label propagation model of Tomasello *et al.* [227] was unable to generate a network matching the empirical ones.

close on the collaborations networks, but also at a small knowledge distance ($\langle d_cos \rangle \approx 0.3$).

Even though the above mentioned procedure for defining communities is promising, we foresee a possible problem. We expect to find that the distribution of the number of scientists at $d_cos < 1$ to have a higher average compared to the distribution of community sizes found in [227]. This should occur because the number of possible PACS (i.e., of knowledge dimensions) is fixed and hence, over the years even really different scientists can get a $d_cos < 1$. We leave the proof of this for future research.

Chapter 7

Modelling scientists' collaborations using knowledge differences

Summary

We propose an agent-based model to reproduce scientists' co-authorship activities using *only* scientists' positions in the knowledge space. Both co-authorship activities and knowledge positions of scientists are obtained using the APS data and the methodology described in the previous chapter. In our model, agents represent scientists, are assigned with a knowledge position and are randomly activated to initiate a collaboration (i.e., a co-authorship activity). When an agent is initiating a collaboration, we use the empirical distribution of pre-collaboration distances to choose the additional collaborators. After selecting these, we form the collaboration between the selected agents and the agent initiating the collaboration. Our simple model reproduces the non-linear trend of scientists' productivity (i.e., the number of co-authored publications per scientist) as a function of the average pre-knowledge distances. Limitations and extensions of the model are discussed.[1]

[1]This chapter contains unpublished work.

7.1 Introduction

Science is done more and more in collaboration and less in isolation. The clearest evidence of this phenomenon can be seen in the increasing number of co-authored publications by scientists. Understanding what brings scientists together for their co-authorship activities is still ongoing research and focus of this chapter. Most previous works have either reproduced these activities considering the underlying social network linking scientists [13, 227, 242] or studied their effect on scientists' productivity and careers [122, 161, 162]. Here, instead, we propose a model that reproduces co-authorship activities by focusing on the different, but complementary expertise of scientists.

A successful perspective used to analyze and reproduce co-authorship activities is based on network analysis [13, 15, 145, 242]. Petersen [161] analyze scientists' ego-network to determine the role of repeated co-authorship activities in the impact of scientists' careers. Ramasco *et al.* [180] provides a self-organizing model to reproduce the collaboration network represented as a bipartite network. Other works instead focused on the interplay between co-authorship activities and the citation network [26, 140, 141]. Additionally, Tomasello *et al.* [227] and Sun *et al.* [219] provide agent-based models for reproducing the collaboration network across scientific fields and the evolution of these fields, respectively. With the exception of [26, 219], all these works neglect the different expertise of scientists when analyzing or defining the determinants behind co-authorship activities.

Different expertise is known to help in producing novel work. In everyday life, this phenomenon goes under the expression that "two heads are better than one". This has been shown to be true in many different scenarios ranging from cognitive experiments involving people Bahrami *et al.* [12] to real-world alliances among firms Baum *et al.* [17], Nooteboom *et al.* [154]. Despite the clear evidence about the importance of this phenomenon, it is often discarded when modeling co-authorship activities among scientists. This chapter aims to fill this research gap by studying and modeling these activities using the different expertise of scientists.

We determine the expertise of scientists by embedding them in a knowledge space and by assigning knowledge positions. To do this, we follow the same procedure as discussed in the previous chapter (see Chap. 6). We restrict our analysis to scientists publishing papers in physics, and we use the APS data (see Chap. 2 and Chap. 6). With this data and the methodology previously introduced, we quantify scientists' expertise and determine their differences. In the model that we are about to propose, these differences will be the determinants of co-authorship activities. In other words, we aim to show that the knowledge positions of scientists are determinants of their collaborations.

The remainder of the chapter is divided into four sections. In Sect. 7.2 we further clarify the aim of our model and in Sect. 7.3 we describe it. In Sect. 7.4 we compare the model with the empirical data by performing a parametric and non-parametric analysis. Finally, in Sect. 7.5, we discuss the results of our model, its limitations, and its possible extensions.

7.2 Aim of the model

In Chap. 5, we have found that knowledge positions of firms are rather determinants than consequences of their collaborations. We now want to verify if a similar statement holds in science. By using bibliographic data to determine co-authorship between scientists, we have only information about *successful collaborations*. With successful collaborations, we mean those collaborations that resulted in a paper accepted in a peer-review journal. Hence, we aim to verify whether knowledge positions of scientists are determinants or consequences of their successful collaborations.

We assume successful collaborations indicate the productivity of scientists. Indeed, the main scope of co-authorship activities is to publish papers. This means that we can quantify scientists' productivity (when collaborating) by looking at the number of their successful collaborations, i.e., of their published papers. In Fig. 6.14 of the previous chapter, we have shown how the productivity of scientists depends on their average knowledge distance from their collaborators. We have found that on average scientists' productivity starts at around three papers for an average knowledge distance equal to 0. It increases

to a maximum at around 12 papers for an average knowledge distance equal to 0.3. Then, it monotonously decreases to two papers for an average knowledge distance between 0.3 and 1.

We propose a new agent-based model that reproduces scientists' productivity as a function of the pre-collaboration distances. In other words, our model will take the empirical pre-collaboration distances as input and will try to reproduce scientists' productivity, i.e., the successful collaborations.

One could argue that we could simply use the label-model used in Chap. 5 able to simulate collaborations and knowledge exchange. Indeed, with this model, we have also reproduced different collaboration networks in science [227]. On the other hand, in this model, agents decide about their partners depending on label attributes representing the membership to circles of influence. While now we aim at studying *only* the role of knowledge in determining scientists' collaborations. This is the reason why we develop a new model instead of using the previous one. Additionally, we are not interested in studying changes in the knowledge positions of scientists. Thus, our new model will neglect possible knowledge exchange occurring among collaborating scientists.

What we propose to do is quite challenging. Indeed, it is clear that the success of a collaboration is not only dependent on the knowledge differences between scientists. Other important factors can be included: the recognition of scientists in their community, cultural differences between scientists, etc. At the same time, if our model can reproduce scientists' productivity, we will indirectly verify that the pre-collaboration distances are proper explanatory variables for predicting co-authorship activities. By this, we would show that knowledge positions can be assumed to be determinants of collaborations.

7.3 Description of the model and simulation procedure

Our agent-based model simulates collaborations occurring from a starting year t until $t + \Delta t$ where $\Delta t = 2$ years. In other words, we consider collaborations occurring during time windows of two years. We choose this value as scientists'

productivity is stable during such time windows (see Figs. E.2 and E.1) and to limit the computational efforts needed for the simulations.

In our simulation, we restrict our attention to 11 non-overlapping time windows between 1984 to 2006. Note that we discard time windows before 1984 as we have less than 1000 established scientists in each of them. While we have discarded the most recent time windows from 2006 to 2010 as the empirical distances matrix $\mathbf{D}_{ik}$ becomes big (> 15 GB). This requires the code to be executed on a cluster with dedicated RAM, and hence, we leave this for future research. At the same time, we firmly believe that the results should not be notably different in these last time windows.

7.3.1 The model

Agents represent established scientists, i.e., scientists that have already participated in collaborations in previous years. This means that we do not consider newcomers, i.e., scientists that co-author their first publication in the time window under analysis. We assign two features to agents: *knowledge position* and *activity*. Knowledge positions represent the expertise of the agents. These positions are taken from empirical knowledge positions of real scientists using the methodology described in Chap. 6. Agents' activities represent scientists' propensity to initiate collaborations. We extract these from empirical data by counting the number of publications that each scientist has co-authored.

The model can be divided in three parts: `Initialization`, `Iteration` and `Termination`. The central part of the model is in the `Iteration` where we simulate scientists' collaborations. This can be divided into four steps: Collaboration size sampling, Agent activation, Selection of Collaborators, and Collaboration formation. While the `Initialization` and `Termination` of the model are more technical parts. Before discussing these technical parts, we describe the core of the model, i.e., the `Iteration`.

Collaboration size sampling. We sample without replacement the size m of one collaboration from the empirical distribution of collaboration sizes. Note that m is the number of established scientists present in the collaboration.

Agent activation. After selecting the size of the collaboration m, we draw an established agent to initiate the collaboration. Each agent is drawn with a probability proportional to her activity. In other words, an agent i is picked to initiate a collaboration with a probability $q_i = a_i / \sum_k a_k$, where a_i is her activity. Recall that agents' activities represent the heterogenous propensities of real scientists to initiate collaborations.

Selection of Collaborators. Once we have selected the initiator and the size m of the collaboration, we select the other $m-1$ collaborators. These are chosen depending on their knowledge distances from the initiator. We sample with replacement $m-1$ distances from the empirical distribution of pre-collaboration distances, and we choose $m-1$ agents at (about) the sampled distances from the initiator.

Collaboration formation. Once we have selected all the collaborators, we create the collaboration. This means that all the collaborators have now a new publication. The information about agents' new publication is used to compute their productivity and their average distance from their collaborators. To understand how we compute these quantities is necessary to understand the more technical parts of the model, i.e., the `Initialization` and the `Termination`. These are described in the next section, where we describe the simulation procedure.

7.3.2 Simulation procedure

Simulation `Initialization`. Given a starting year t, we create N agents representing the N established scientists collaborating between t and $t + \Delta t$. We initialize the N agents by assigning empirical knowledge positions (computed up to year t) and empirical activities (computed between $t + \Delta t$). Also, we initialize three vectors of length N with all zeros. The first vector $\mathcal{P}$ will contain the productivity of the agents, i.e., their number of publications. The second vector $\mathcal{L}$ will contain the number of collaborators (with repetition) for each agent. The third vector $\mathcal{C}$ will contain the average distance of an agent from her collaborators.

Simulation `Iteration` and update. After initializing the agents and creating the $\mathcal{P}$, $\mathcal{L}$ and $\mathcal{C}$ vectors, we simulate the collaborations using the four steps described in Sect. 7.3.1. This means that we iteratively sample a collaboration size m, activate an agent i, choose her collaborators C, and create the collaboration. When we create the collaboration, we update the three vectors $\mathcal{P}$, $\mathcal{L}$ and $\mathcal{C}$. First, we update the components of the productivity vector relative to the collaborating agents, $\forall i \in C \ \mathcal{P}_i \leftarrow \mathcal{P}_i + 1$. Second, we update the number of collaboration links of each collaborating agent by $(m-1)$, $\forall i \in C \ \mathcal{L}_i \leftarrow \mathcal{L}_i + (m-1)$. Third, we update the aggregate knowledge distance of each agent by adding her distances from her collaborators, $\forall i, j \in C \ \mathcal{C}_i \leftarrow \mathcal{C}_i + D_{ij}$ where D_{ij} is the knowledge distance between i and j.

Simulation `Termination`. We keep sampling without replacement from the empirical distribution of collaboration sizes until we have sampled all the collaborations. Once we have finished simulating the collaborations, we store for each agent her productivity and the average knowledge distance from her collaborators.

For each time window with different starting year t, we simulate 20 times scientist collaborations with our model. We summarize the model using a pseudo-code in Algorithm 1 and provide its input and output in Table 7.1.

7.4 Results

Our model aims to show that scientists' knowledge positions are determinants of collaborations. We do this by showing that the pre-collaboration distances, i.e., differences in knowledge positions, can be used to reproduce the productivity of a scientist p_i in function of her average knowledge distance $\langle \Delta x_{ij} \rangle_{\mathcal{N}_i}$ from her collaborators $\mathcal{N}_i$. In Figure 7.1(a), we compare the distribution of p_i coming from simulation and from the empirical data for the time window starting in 1988 and ending in 1990. From this plot, we note that the empirical data and simulated one overlap quite a lot. However, it is not possible to determine how much the model is consistent with the observation. To analyze the similarity and differences between the simulated and empirical data, we

Input	
M	Collaboration sizes
A	Scientists' activity
E	Pre-collaboration distances
$\mathbf{D}$	Scientists' distance matrix

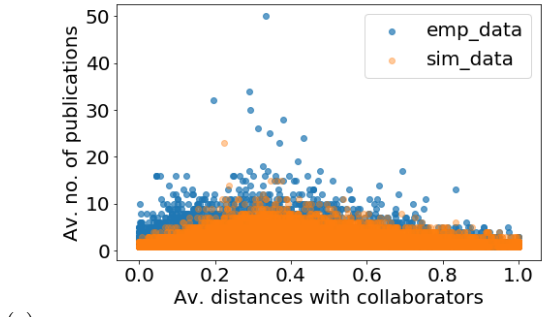
⟹

Output	
$\mathcal{P}$	Productivity
$\mathcal{C}/\mathcal{L}$	Average collaboration distances

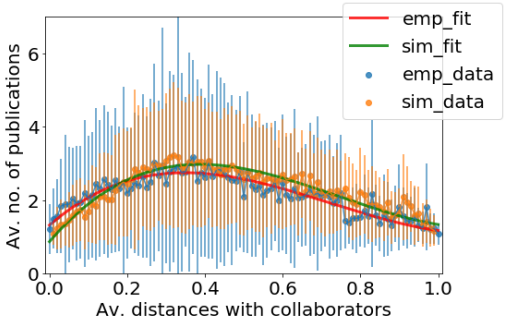
Table 7.1: Summary of the input and the output of the model.

Data: M, A, E, $\mathbf{D}$
Initialization: $N \longleftarrow len(A)$, $r_{indices} \in \mathcal{R}^N$, $\mathcal{P} \in \mathcal{R}^N$, $\mathcal{L} \in \mathcal{R}^N$, $\mathcal{C} \in \mathcal{R}^N$;
$r_{indices} \longleftarrow (1, 2, \ldots, N)$, $\mathcal{P} \longleftarrow \vec{0}$, $\mathcal{L} \longleftarrow \vec{0}$, $\mathcal{C} \longleftarrow \vec{0}$;
while $M \neq \emptyset$ **do**
  $r_{indices} \longleftarrow$ random shuffle $r_{indices}$;
  for $i \in r_{indices}$ **do**
    $p \longleftarrow x \in \mathcal{U}[0:1]$;
    if $p < A_i$ **then**
      $m \longleftarrow$ sample m from M; ▷ Agent-i is actived
      $C \longleftarrow \{i\}$; ▷ Create set of collaborators
      for $l = 1$ **to** $m - 1$ **do**
        $d \longleftarrow$ sample d from E;
        $j \longleftarrow \arg\min_{k \in N/c} |d - \mathbf{D}_{ik}|$;
        $C \longleftarrow C \cup \{j\}$; ▷ Add j to set of collaborators
      end
      for $l \in C$ **do**
        $\mathcal{P}_l \longleftarrow \mathcal{P}_l + 1$; ▷ Update productivity,
        $\mathcal{L}_l \longleftarrow \mathcal{L}_l + (m - 1)$; ▷ Update number of links and
        for $l' \neq l \in C$ **do**
          $\mathcal{C}_l \longleftarrow \mathbf{D}_{ll'}$; ▷ Update distances from collaborators
          $l'++$;
        end
      end
    end
    $M \longleftarrow M/m$
  end
end
Save output: store $\mathcal{P}$ and $\mathcal{C}/\mathcal{L}$

Algorithm 1: Pseudocode. The input data are the lists containing the collaboration sizes M, the activities of the agents A, the pre-collaboration distances E and the matrix containing the empirical distances between the agents $\mathbf{D}$.



(a)



(b)

Figure 7.1: The empirical and simulated distributions of scientists' productivity in function of the average knowledge distance from their collaborators in orange and blue, respectively. In (a), we provide this as scatter plot where each point represent a scientist. In (b), we provide the fit of the data using Eq. (7.1).

follow two approaches. First, we perform a parametric fit. This allows to gain insight into common or different trends between simulated and observed productivity. Second, we perform a non-parametric analysis to compare the obtained distributions of data points. In particular, we use the non-parametric test for multivariate distributions introduced by [14].

7.4.1 Parametric analysis

In order to perform a parametric analysis of scientists' productivity, we propose a fitting function. To obtain this, we now specify the productivity p_i, of an individual scientist i given her collaborators $\mathcal{N}_i$. Because we focus on the impact of knowledge in collaborations, we consider the knowledge distances Δx_{ij} as the only influence that is reflected in p_i. Specifically, we omit effects resulting from the shared workload to write a publication or from the prestige of the co-authors. In other words, our central assumption is that the productivity p_i is dependent only on one types of variables: the knowledge distances between the focal scientist i and her collaborators $\mathcal{N}_i$.

We propose the following fitting function for the productivity:

$$p_i = a + b \, \langle \Delta x_{ij} \rangle_{\mathcal{N}_i} + c \, \langle \Delta x_{ij} \rangle^2_{\mathcal{N}_i} + d \, \langle \Delta x_{ij} \rangle^3_{\mathcal{N}_i} \tag{7.1}$$

where $\mathcal{N}_i$ are the collaborators of i, Δx_{ij} is the pre-collaboration distance between scientists i and j, $\langle \cdot \rangle_{\mathcal{N}_i}$ is the average taken considering the $\mathcal{N}_i$ collaborators, a, b, c and d are the parameters that need to be fitted. To interpret the different terms in the proposed fitting function p_i, we have to consider that pre-collaboration (knowledge) distances give rise to both *benefits* and *costs*.

Benefits and costs. We expect a and b to be positive as the sum $a + b \, \langle \Delta x_{ij} \rangle_{\mathcal{N}_i}$ quantifies scientists' *benefits* from collaborations. In particular, the first benefit term a reflects the empirical finding that a scientist i co-authors papers also with collaborators at a knowledge distance $\Delta x_{ij} = 0$. The second benefit term $b \, \langle \Delta x_{ij} \rangle_{\mathcal{N}_i}$ is linear in the knowledge distance, and it mimics the linear relation between cognitive distance and novelty [154]. Indeed, by using the words of [154], "*When people with different knowledge and perspectives interact, they stimulate and help each other to stretch their knowledge for the purpose of bridging and connecting diverse knowledge*". At the same time, knowledge differences also determine a *cost* in collaborations. If knowledge differences become too big, then communication between scientists also becomes very difficult. Indeed, it needs a lot of effort before one can understand all the comments, ideas, and contributions of someone with a completely different

scientific background, even if this person is a renowned scientist. To capture this, in Eq. (7.1) we have a third quadratic term $c \langle \Delta x_{ij} \rangle^2_{\mathcal{N}_i}$ that overcomes the linear benefit after an optimal knowledge distance. We expect c to be negative as it is a cost, and it should decrease productivity.

In Eq. (7.1), we have also added a final cubic term ($d \langle \Delta x_{ij} \rangle^3_{\mathcal{N}_i}$) for two reasons. First, there is no reason for imposing the productivity to be symmetric with respect to the knowledge distance. Second, the minimum value that productivity can assume is 1, and hence, we need a positive cubic term that allows compensating the quadratic term. For this reason, we expect $d > 0$.

In Fig. 7.1(b), we report the simulated and empirical productivity fitted using Eq. (7.1) for the time window starting in 1988 and ending in 1990. We perform the fit using the `curve_fit` function in the `scipy` package[2]. Again from a visual inspection, we see that the simulation well matches the empirical data. Precisely the fitted productivity functions, coming from simulations and empirical data, have the same non-linear trend.

Typically, a model provides an expectation that it is compared with several empirical observations. While in our case, we have one observed productivity function (for each time window) and obtain (slightly) different productivity functions from each simulation of our model. For this reason, we have to determine the average simulated parameters and confidence bounds of our model. With these, we can check if the empirical observation is compatible with our model.

To obtain average simulated parameters and confidence bound for each time window, we perform 20 simulations. On each simulation, we fit the productivity function given in Eq. (7.1) and obtain the four parameters, a, b, c and d. We then compute the average of the obtained parameters across simulations. Similarly, we can obtain confidence bounds on these parameters by computing their empirical errors across simulations. By assuming that the fitted parameters are normally distributed, we can assume that 95% of the simulated parameters are contained in two standard deviations from their average.

[2]This function solves non-linear least squares problem `https://scipy.org/`. We use the default algorithm, which is the Levenberg–Marquardt algorithm [118, 131]

Parameters	Empirical	Simulated
offset, a	1.321	0.980 ± 0.035
benefits, b	9.15	8.54 ± 0.44
costs, c	-17.3	-12.3 ± 1.2
saturation, d	7.99	9.64 ± 0.83

Table 7.2: Fitted parameters for the productivity function between 1988 and 1990. From right to left, the parameters coming from fitting the empirical and simulated data. The empirical parameters are the averages obtained from fitting each simulation, separately. We also report their standard error (multiplied by two). For all parameters, we truncate the digits at the second significant digit of the standard errors coming from the simulated data.

Note that given the high dispersion of the data it is not worth to look at the error of the estimated parameters directly from the fit. These errors are small even though the fitted model cannot explain the variance of the data, and they should not be used. In Table 7.2, we report the parameters estimated from simulated and empirical data for the time window starting in 1988 and ending in 1990. We verify that the fitted productivity from simulations and empirical data have a similar trend as they have similar parameters. They both have a positive offset a matching the fact that there collaborations among scientists with identical knowledge. Additionally, we find that the quadratic terms are negative ($c < 0$), i.e., they are costs, while the linear terms are positive ($b > 0$), i.e., they are benefits. Last, but not least, we find that at large values of knowledge distance both empirical and simulated data have positive cubic terms ($d > 0$) to balance the negative quadratic terms.

Even though our simulated productivity well matches the non-linear trend of the empirical one, we also find differences. The most significant difference between the two curves is at small knowledge distances. In our simulation, we obtain fewer scientists that are productive with collaborators at $\langle \Delta x \rangle < 0.2$. We provide an explanation of this in the Sect. 7.5. Moreover, we find that the fitted parameters obtained from the empirical data are not contained in two standard errors coming from the simulations (see Table 7.2). This hints at us that there might be statistically significant deviations between data and simulations. To verify this, we perform a statistical test in the next section.

Observed statistic	Critical value ($\alpha = 95\%$)	H_0 (p-value)
37.238	2.459	REJECTED ($< 10^{-5}$)

Table 7.3: Results of the Cramer test coming from comparing the empirical and the simulated distribution of (*productivity*, *average collaboration distances*) pairs.

7.4.2 Non-parametric analysis

We perform a non-parametric statistical test to understand how well our simulations reproduce the observed data. To do this, we consider simulations and empirical data as bi-dimensional distributions of random vectors. Each random vector represents a scientist where the first variate is her productivity, while the second variate her average knowledge distance. To compare the bi-dimensional distributions of empirical and simulated random vectors, we use the multivariate two-sample test of Baringhaus and Franz [14], also known as the *Cramer test*.

The Cramer test is based on a simple idea. Given two samples of points $\{\vec{X}_i\}$ and $\{\vec{Y}_i\}$ of size m and n, we can compute three average distances: i) the average distance between points in the first sample $\langle d \rangle_x$, ii) the average distance between points in the second sample $\langle d \rangle_y$, and iii) the average distance between points across the two samples $\langle d \rangle_{(x,y)}$. Then $\frac{1}{2}(\langle d \rangle_x + \langle d \rangle_y) \geq \langle d \rangle_{(x,y)}$ where the equality holds when both samples of points belong to the same distribution [138]. To use this idea, the Cramer test computes the Cramer statistic that is $T_{mn} = \frac{mn}{m+n}(2\langle d \rangle_{(x,y)} - \langle d \rangle_x - \langle d \rangle_y))$ and we compare this to a critical threshold T_α. The critical values T_α is the upper α-quantile coming from the empirical distributions of distances, and it allows to reject or accept the following hypothesis with a confidence bound α:

H_0: The two samples of points belong to the same distributions.

H_1: The two samples of points belong to different distributions.

When $T_{mn} < T_\alpha$, we *can* reject H_1 with confidence α (and can accept H_0), i.e. with probability α the two sample of points belong to the same distributions. While, when $T_{mn} \geq T_\alpha$, we *cannot* reject H_1 with confidence α (and cannot

accept H_0), i.e. with probability α the two sample of points belong to different distributions.

In Table 7.3, we report the results from the test. We use the implementation of the test provide in the `R` package `cramer`. Note that given the large sample size, we used the eigenvalue estimation, instead of the Monte-Carlo bootstrap method of the package. We find that we cannot reject H_1, i.e., the hypothesis that the empirical and simulated data come from different distributions with 95% of confidence. This means that even though we find a good match in the parametric analysis, the distributions are still statistically different from each other. We believe the most significant mismatch is in two differences between the empirical and the simulated of the data. First, in the empirical data, all scientists have at least participated in one collaboration, while in the simulations, an agent has productivity equal to zero if she is never sampled. Second, the empirical data contains some outliers, i.e., scientists with much higher productivity, that are not reproduced in the simulations. We further discuss the limitation of our model in Sect. 7.5.

Note that the information reported in Table 7.3 comes from on single simulation of collaborations for the time window starting in 1988 and ending in 1990. When comparing all the simulations with the empirical data, we always have to reject the hypothesis that simulations and empirical data come from different distributions with 95% of confidence. This means that with probability $p = 0.05$, the test is giving a wrong result, i.e., with probability 0.05, the test rejected H_0 even though it was correct. On 20 trials, the probability that the test fails more than two times is $\sum_{i=3}^{20} \binom{20}{i} p^i (1-p)^{20-i} = 1 - \sum_{i=0}^{2} \binom{20}{i} p^i (1-p)^{20-i} = 0.075$. This means that with a probability higher than 0.9, the test will give correct results for 18 out of the 20 trials. In other words, with a probability higher than 0.9 the test correctly rejects H_0 $18/20 \approx 90\%$ of the time. By this, we can say that our model has statistically significant deviations from the empirical data.

7.5 Discussion and outlooks

We have developed an agent-based model that allows capturing the productivity of scientists as a function of their positions in a knowledge space. Our model is an activity-driven model [160], meaning that agents' propensity to initiate a collaboration is drawn from empirical data. Additionally, we use empirical information on the number and the size of collaborations to determine how many and how large the simulated collaborations should be, like in [224, 233]. The novel aspect of our model is that we use knowledge positions of agents to determine their collaborators. To do this, we assign agents with knowledge positions using scientists' empirical knowledge positions, computed in Chap. 6.

From a parametric analysis, we find that our simulations reproduce the functional form of the observed scientists' productivity. Precisely, we have proposed a scientist's productivity function based on knowledge distances between scientists that allows us to fit the observed productivity function (see Eq. (7.1)). This is a cubic function with four parameters weighting constant and linear benefits, quadratic costs, and a cubic saturation in the knowledge distances. We find that the fitted parameters from simulated and empirical data are close and reproduce the same non-linear trend for the productivity function.

From a non-parametric analysis, we find that the simulated distribution of (*productivity*, *average collaboration distances*) pairs are statistically different from the empirical one. We obtain this result by comparing the bi-dimensional distributions using the Cramer test, a non-parametric test [14]. We argue that the statistical difference is due to two main factors. *First*, the empirical data contains some extremely productive scientists not reproduced in the simulations (see Fig. 7.1). We believe that the higher productivity of these scientists is not dependent on their knowledge positions. Indeed other factors, like the visibility of the scientists [13] or the availability of findings [126], could be important determinants for collaborations. Recall that we have discarded these factors in order to focus only on the role of knowledge in determining collaborations. *Second*, the empirical distribution contains a higher number of scientists that have smaller average distances compared to the simulated one (again, see Fig. 7.1). This occurs as in our model, we sample an initiator and

pick the remaining collaborators by *only* considering their distance from the sampled initiator. This implies that with high probability, the picked collaborators have a significant knowledge distance among them. Recall that when checking the knowledge distance among random pairs of scientists, we have obtained a distribution of knowledge distances with one prominent peak at high knowledge distance (see Fig. 6.4(c)).

To solve the above problems, we consider two possible extensions to our model. To reproduce the presence of scientists with extremely high productivity, we could introduce a bias when picking collaborators. We should favor those that have already been participating in many collaborations. Such an approach has shown to be successful in the activity-driven model of Tomasello *et al.* [224], also used in Chap. 5. While to obtain more scientists with smaller average distances from their collaborators, we could define a collaboration fitness that considers the knowledge distances among *all* collaborators. We could compute and assign such fitness to each simulated collaboration by modifying the productivity function defined in the parametric analysis. Then a collaboration is created if its fitness is higher than a given threshold. Note that a similar approach was used to successfully reproduce the size distribution of collaborations in R&D alliances by Tomasello *et al.* [222].

We have performed some exploratory work on both the extensions mentioned above. We have found that the challenge on these extensions is balancing between the effect of the utility function of favoring a specific knowledge distance and the high dispersion of the (observed) data. Further research needs to be done to determine the effectiveness of the proposed extensions.

Finally, let us comment on the similarities and the differences between this model and the one presented in Chap. 5. Both models reproduce collaboration activities and take the number and the size of collaborations as input. However, the considered collaborations occur in really different domains: science and industry. Both models are activity-driven, meaning that agents are assigned with an activity attribute that captures their propensity to start collaborations. Moreover, in both models, we assign knowledge positions to agents as second attributes. However in Chap. 5, knowledge positions change over time, and they do not play a role when choosing the collaborators. Instead,

in this chapter, knowledge positions are constant attributes of the agents, and they are the *only* determinant of the collaborations.

Note that the main determinant of the collaborations in Chap. 5 were the (constant) label attributes assigned to agents. These labels capture the membership of the agents to well-defined circles of influence [224]. Interestingly enough, the label approach used in Chap. 5 was also successfully used to reproduce the co-authorship activities among scientists [227]. This motivates us to link labels and knowledge positions of the agents (see Sect. 6.7). At the same time, it is not trivial how to create this link, and more research on this possible connection is still needed.

To conclude, in the present chapter, we have developed a simple, yet efficient agent-based model able to reproduce the productivity of scientists as a function of their knowledge positions. By proposing a theoretical productivity function, we verified that the model reproduces the functional form of the observed productivity, but not all its details. With a non-parametric analysis, we have identified a statistical incongruence between the simulated and observed scientists' productivity. We have argued about the origin of the incongruence, and we have proposed a more refined agent-based model to cure for the identified incongruence. By all this, we shed new light on the role of different knowledge among scientists in determining their co-authorship activities.

Chapter 8

Scientists' mobility: An empirical analysis

Summary

This chapter makes two essential contributions to understand the mobility patterns of scientists. First, by combining two large-scale data sets, we are able to reveal the *geographical* career trajectories of scientists. Each trajectory contains, on the individual level, information about the cities and the time spent there. A statistical analysis gives empirical insights into (i) the geographical distance scientists move in order to obtain a new affiliation and (ii) scientists' age when moving. From the individual career trajectory, we further reconstruct the world mobility network of scientists, where nodes represent cities and links in- and outflow of scientists. We analyze the topological properties of this network with respect to degree, local clustering coefficient, path length, and neighbor connectivity. The second important contribution is an analysis of the temporal correlations of scientists' career trajectories. This analysis is performed at the country, and by this, we verify whether international corridors favor the mobility of scientists. Moreover, we perform the analysis of temporal correlation at the affiliation level, and hence, we test whether it is adequate to model scientists movements across institutes using a network perspective.[1]

[1]Based on [231]

8.1 Introduction

So far, we have quantified knowledge and proxied its exchange among scientists and firms in terms of patents and scientific publications. These are knowledge artifacts capturing one dimension of knowledge, namely explicit knowledge [96, 151, 166]. However, knowledge has also another dimension, often called either tacit or implicit knowledge. To study and quantify this other dimension of knowledge is difficult by definition as it is the knowledge that cannot be encoded, and humans can exchange only via shared practice [96, 151, 166].

In this chapter, we address the problem of studying tacit knowledge by analyzing scientists' career trajectories. Indeed, scientists diffuse explicit knowledge by publishing their research and tacit knowledge by physically moving across different locations. Thus, by studying scientists' career trajectories, we can indirectly study the diffusion patterns of tacit knowledge. We provide an empirical analysis of scientists' career trajectories at three different levels: country, city, and affiliation level.

In Sect. 8.2, we analyze the geographical properties and timing of scientists' movements at the city level. To do this, we use the largest open-access bibliographic data set on life sciences, `MEDLINE` to reconstruct the geographical trajectories of scientists (see Chap. 2 for details). Then, by taking a network perspective, we reconstruct the global mobility network of scientists. On this, we compute the distributions of four network measures to determine the geographical properties of scientists' career trajectories.

In Sect. 8.3, we perform an analysis of the temporal correlations of scientists' career trajectories at the country and affiliation level. To do this temporal analysis while retaining a network perspective, we use the mathematical notion of paths and multi-order graphical models [201]. At the country level, we investigate whether there are shared international corridors in the movements of scientists. At the affiliation level, we analyze whether scientists' careers are dependent only on their current working institutions or also on the previous ones.

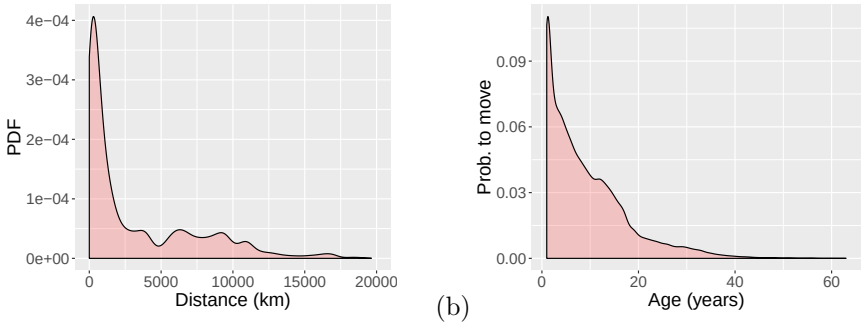


Figure 8.1: (a) Distribution of movement distances of scientists. (b) Distribution of movements dependent on the (academic) age of scientists.

We conclude with Sect. 8.4 by summarizing and interpreting our empirical findings. Note that the main contribution of this chapter is two-folded. First, we provide an analysis of scientists' career trajectories at the country, city, and affiliation levels. This analysis allows us to develop a broad understanding of the properties of these trajectories. Second, our analysis uses state-of-the-art network-analytic methods to study temporal aspects while retaining a network perspective. By this, we are able to uncover new patterns of scientists' career trajectories.

8.2 Individual and global mobility of scientists

By using the MEDLINE data and the method presented in Sect. 2.1.1F, we reconstruct the career trajectories of individual scientists at the city level. Using these trajectories, we analyze the geographical properties of the movements of individual scientists. Then, by taking a network perspective, we analyze the global mobility of scientists as a system.

8.2.1 Statistics of geographical career trajectories

The information about the sequence of cities a scientist was based during her career allows us to analyze the *distance* she moved when changing her

affiliation. We use the Haversine formula to compute the geodesic distance between the geo-locations of the respective cities, measured in kilometers. The distribution obtained from 62 465 scientists moving between 2000 and 2008 is shown in Figure 8.1(a). We note that it is a left-skew distribution with a median of 1 000 km, i.e., most scientists find a new affiliation in cities within a radius of 1 000 km around their current affiliation. However, movements of more than 6 000 km toward distant cities are also quite frequent.

The data also allows us to relate the frequency of such moves to the age of scientists. Because the physical age of scientists is not recorded, we have to rely on their *academic age*, t_i^a, also measured in years. $t_i^a = 0$ when the scientist publishes her first paper, according to our database (which is probably a physical age of about 25 years). The frequency of any recorded move irrespective of the distance over the academic age t^a is shown in Figure 8.1(b). Again, it is a left-skew distribution with a median of 7 years. The properties fo this distribution the known fact that the mobility of scientists drastically decreases with age [35, 237]. However, we find frequent moves even at the (physical) age of retirement.

8.2.2 Reconstructing the mobility network of scientists

While the career trajectories and their statistics refer to individual scientists, we can also analyze the network that results from *aggregating* all of the career trajectories of a given year. This second analysis moves the discussion of movements between *cities* to the macro level. For each year, we calculate the *number of scientists* $N_K(t)$ in a given city K from their publications, taking unique geo-located authors into account. We further calculate for each year t the number of scientists $\Delta N_{K\leftarrow L}(t)$ moving *into* city K from another city L, i.e. the *inflow*, and the number of scientists $\Delta N_{L\leftarrow K}(t)$ moving out of city K to another city L, i.e. the *outflow*.

Figure 8.2(a,b) show the respective distributions for the aggregated inflow $\Delta N_K^{\text{in}}(t) = \sum_L \Delta N_{K\leftarrow L}(t)$ of scientists into city K and the aggregated outflow $\Delta N_K^{\text{out}}(t) = \sum_L \Delta N_{L\leftarrow K}$ of scientists out of city K. The aggregate inflow and the outflow are computed during three different time windows centered in 2000, 2002 and 2004, meaning that each city is considered three times (once for every

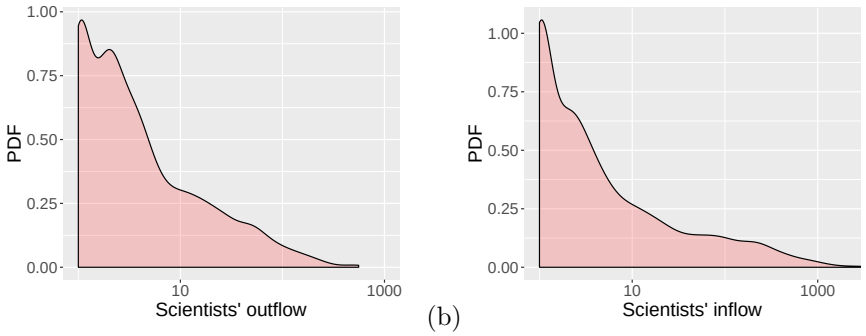


Figure 8.2: Distributions of (a) inflow of scientists into any city, (b) outflow of scientists from any city. The x-axis is in log-scale.

time window). Again, we note the left-skew distribution for both quantities, which indicates the *heterogeneous* contribution of cities to the global movement of scientists.

For any given pair (K, L) of cities we can then calculate the *total flow* of scientists between these two cities. This is the total number of scientists exchanged between K and L, $\Delta N_{L \leftarrow K} + \Delta N_{K \leftarrow L}$. The total flow allows us to visualize the mobility network of scientists a the world level, as it is shown in Figure 8.3. The links are undirected, but weighted according to the total flow.

Fitness of a city. The calculated inflow and outflow already makes clear that cities are very different with respect to their attractiveness for scientists. Obviously, a small number of cities are more attractive, which can be explained to a large extent by the reputation of the academic institutions hosted there. Hence, it makes sense to assign to each city $K \in M$ a *fitness value* $F_K(t)$ reflecting the quality of their academic institutions. This fitness value is not precisely known, but can be estimated from available data, for example taking different *university rankings* into account. We will not describe in detail how we measure the city fitness, suffice it to say that we measure it through a citation weighted output metric (see Chap. 2 for details). We assume that such a measure reflects the scientific attractiveness that scientists associate with cities. The actual values are not relevant since we are primarily interested in

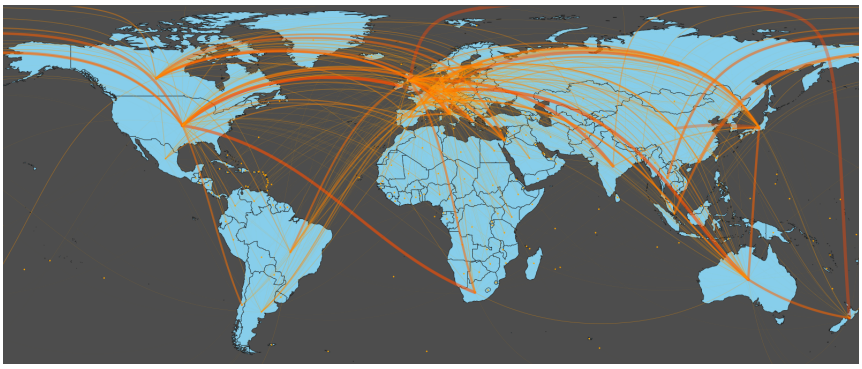


Figure 8.3: The mobility network of scientists in between 1990 and 2008. The link width and the color indicate the magnitude of the *total flow* between any two cities. For visualization purpose, the total flows have been aggregated at country level and logarithmically scaled.

the *ranking* of cities resulting from this measure, i.e., in the fitness *relative* to the others. We note that city fitness can change over time. For an extensive empirical investigation about city fitness and its interplay with scientists' career trajectories, see [237].

8.2.3 Topological properties of the mobility network

In order to further characterize the mobility network utilizing topological properties, we aggregate the mobility networks for the period of time 2000-2008. On this *aggregated network*, we calculate standard measures that are common in network analysis. These measures include the *degree distribution* $P(d)$, where d is the number of cities scientists in a given city either move to or come from. Already Figure 8.3 indicates that this is a very broad distribution. Some cities act as hubs and have a large degree. However, most cities only have a small degree. This result is confirmed by the degree distribution shown in Figure 8.4(a).

The distribution of *path lengths*, shown in Figure 8.4(b), measures how many steps are needed to reach, on the network, any city from a given starting

point. The small number of hops indicates that the network is very dense in a *topological sense*, not necessarily in a geographical one.

The *local clustering coefficient*, on the other hand, measures whether three neighboring cities (with respect to their network distance) form closed triangles, i.e., whether there is an exchange of scientists between them. Figure 8.4(c) shows the distributions of these values, and we find that most cities have a small local clustering coefficient.

The *neighbor connectivity* measures to what extent cities with a certain degree are connected to other cities with a similar degree. Figure 8.4(d) shows a non-monotonous dependency. Cities with a low degree tend to show an *assortative* pattern, i.e., they are connected to cities that have a similar number of neighbors. Cities with a high degree, which are characterized as *hubs* above, are rather connected to cities with a lower degree, i.e., they are *disassortative*. This gives us already on the topological level important information about the origin of scientists coming to the hubs and the destination of scientists leaving the hubs. Obviously, they do not hop between hubs - which would have been indicated by a clearly assortative pattern for hubs.

8.3 Memory effects in scientists' mobility

We now turn our attention to a different aspect of our data. We study whether career trajectories of scientists contain temporal correlations. We start our analysis at the country level, and hence, career trajectories are sequences of countries where scientists move in order to work. Temporal correlations in these sequences would indicate the presence of international corridors (made by more than two countries) that channel scientist movements.

On the one hand, we expect that analyzing scientist trajectories at the country level should give reliable results. By aggregating data at the country level, we have a quite high number of trajectories between many sequences of countries, and hence, we obtain better statistics. On the other hand, aggregating and projecting sequential data might distort the modeling of the data and destroy its temporal properties (see Chap 4 and [204]). For this reason, as the second step of our analysis of temporal correlations, we analyze career trajectories of

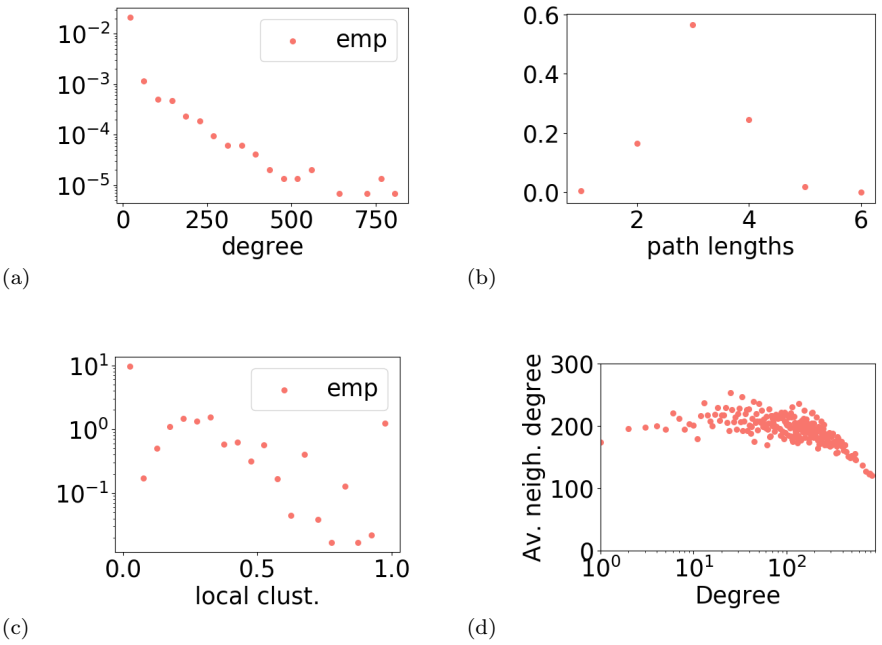


Figure 8.4: Distributions of (a) degrees, (b) path lengths and (c) local clustering coefficients. In (d) we plot the the average degree of neighbors of a node with degree k in function of k.

scientists at affiliation level. By affiliation level, we mean that scientists' career trajectories are sequences of research institutions (affiliations) where they have published a paper. With this, we analyze scientists' careers at a much more fine-grained level compared to the city or country level.

For our analysis at the country and affiliation level, we adopt a path perspective similar to the one used in Chap. 4. In other words, we use the multi-order graphical model developed by Scholtes [201] to model the career trajectories, and then, we will look for the optimal order.

# of countries	# of links (unique)	# of trajectories	[Min, Max]
215	6 913	3 740 187	[1, 32]

Table 8.1: Key statistics of the career trajectories at the country level.

# of universities	# of links (unique)	# of trajectories	[Min, Max]
81	6 340	2 312 376	[1, 42]

Table 8.2: Key statistics of the career trajectories at the affiliation level.

8.3.1 Temporal correlations at country level

We restrict our attention to 3 740 187 individual scientist trajectories across 215 countries between 1990 and 2009. The longest trajectory is of length 32, and 89% of the trajectories have length 1, meaning that we observe most scientists moving only once. This points out that the repeated long term movements are relatively rare, but still many in absolute terms (411 258).

The most frequent trajectories of length one are the ones between UK and USA (30 406), Japan and USA (28 013), USA and UK(26 660). This result is directly dependent on the fact that the USA is actually the largest country made of many states, and for this reason, we always observe in the most frequent trajectories. If we consider only those trajectories not coming or going to the USA, we find that the most frequent trajectories of length one are across the UK and Australia (6 386), Germany and the UK (6 015), and the UK and Germany (5 441).

When considering trajectories of length two, the most frequent ones are between (Japan, USA, Japan) (14 362), (USA, UK, USA)(11 416), and (UK, USA, UK)(10 759). Again, the USA is always present, and we argue this depends on the fact that the USA is the largest country in the data. If we consider only those trajectories of length two that do not go through the USA, we find that the most frequent ones are (UK, Australia, UK)(2 027), (France, UK, France)(2 101), and (Germany, UK, Germany)(2 108). The origin and end of these trajectories are in the same country, and hence, this suggests the presence of a"go back home" phenomenon. In other words, scientists often go back to their (academic) country of origin after working in a different one.

In order to identify temporal correlation in our trajectories, we use the test developed by Scholtes [201] to detect the optimal order for multi-order graphical models. By using `Pathpy` implementation of this test, we find $K_{opt} = 1$ to represent scientists' career trajectories at the country level. This means that there are no statistically significant memory effects in the career trajectories to motivate the use of higher-order network models. In other words, we obtain $K_{opt} = 1$ because the number of empirical trajectories of length 2 or higher is too low compared to the degrees of freedom necessary to represent the data as a second-order network. Note that by analyzing the temporal correlation at the city level, we also find $K_{opt} = 1$.

8.3.2 Temporal correlations at affiliation level

For the analysis at affiliation level, we use the MAG data set (see Chap.2), instead of the `MEDLINE` data used in the previous sections. Note that we had to use `MEDLINE` data as scientists needed to be geo-localized at either city or country level. While now we are not interested anymore in scientists' geographical locations, but only in their affiliations. For this reason, we prefer the MAG data set that contains scientists publishing in more disciplines compared to ones listed in `MEDLINE`.

We restrict our analysis to the top 100 universities in computer science reported in [38]. To do this, we matched the names of the universities using basic string processing and obtained 2 312 376 scientists trajectories of scientists moving across 81 universities. For more details about the universities under analysis, see Appendix D.

By focusing on a specific set of universities, we have to consider two facts when interpreting our results. First, the analysis of temporal correlation will be valid only for the subset of universities analyzed. Second, by analyzing a subset of *top* universities, we are focusing our attention on a particular sub-population of scientists. We are focussing on scientists that have a career in more visible and established institutions. On the one hand, this is a second limitation as we have a biased sample of scientists. On the other hand, we obtain insight into the faculty hiring system of top universities. By this, we will complement and extend the network analysis that Clauset *et al.* [38] have performed.

Among the top analyzed universities, the most frequent trajectories of length one are between Washington University in St. Louis and the University of Washington, between Kyoto University and the University of Tokyo, and between Tsinghua University and Peking University. So a large number of scientists' movements of length one are occurring at the national level, and we find this to be true also when looking at longer trajectories. For example, among the most frequent trajectories of length two, we find trajectories like (Washington University in St. Louis, University of Washington, Washington University in St. Louis), (University of Tokyo, Kyoto University, University of Tokyo). All this hints us that by looking at scientists' career trajectories at the country level, one discards extremely frequent movements that could contain temporal correlations. To check for this, we the test of [201] provided in `Pathpy` and find $K_{opt} = 2$. This means that there are statistical significant memory effects in the data to justify the use of a multi-order graphical model of order two. We discuss this result in the next section.

8.4 Conclusions

This chapter contains important results from an empirical point of view. We have analyzed career trajectories of scientists at the country, city, and affiliation level by using two large scale bibliographic data sets: `MEDLINE` [228, 229] and `MAG16` [213]. For the analysis at the city and country level, we have used the `MEDLINE` dataset as it provides geo-localized information on scientists working in life sciences. While for the analysis at affiliation level, we use the `MAG` as it covers a broader range of disciplines compared to `MEDLINE` data.

When analyzing scientists' career trajectories at city level, we have started by computing the distributions of the number of scientists moving to and from cities. We have found that both distributions are broad and show large variations in the attractiveness of different cities from scientists' perspective. Then, we have reconstructed the global mobility network of scientists. On this network, we have calculated the distributions of four quantities: degree, shortest path length, local clustering coefficients, and neighbor connectivity. We have found that the degree distribution is extremely broad, showing that

few cities act as hubs, while the majority have a small degree. This result confirms that city attractiveness for scientists varies a lot. On the contrary, the shortest path length distribution is narrow and peaked at 3, showing that the network is very dense in a topological sense. The local clustering coefficients distribution is peaked at low values telling us that most cities are not connected with closed triangles. Hence, scientists often move to (away from) hubs from (to) smaller cities, but these smaller cities are not connected. At the same time, from the distribution of neighbor connectivity, we have found a clear disassortative preference for nodes with high degrees. This indicates that hubs are not well connected with each other.

When analyzing scientists' career trajectories at the country and affiliation level, we have determined whether there are temporal correlations in the sequences of countries or institutions where scientists move to work. By performing our temporal analysis at two different levels, we have studied two different aspects of scientists' career trajectories. At the country level, we find no statistically significant temporal correlations in the sequences of countries. This indicates that there are no international corridors that channel the scientists' circulation. By this, we also find that a network perspective is sufficient to analyze the scientists' career trajectories across countries.

At affiliation level, we have analyzed scientists' career trajectories across 80 top universities and found temporal correlations in their careers. These temporal correlations mean that scientists work in universities following specific sequences during their careers. On the one hand, we complement the findings of Clauset *et al.* [38], which determined the hierarchical social structure of universities using a network perspective. On the other hand, we find that a network perspective is an incomplete representation of scientists' career trajectories. To correctly represent these trajectories, we should use a second-order network at the university level. This result might have implications when using network measures to determine the hierarchical social structure of universities or in general, to create university rankings. We leave the analysis of such implications for future research.

In the present chapter, we have performed an extensive analysis of scientists' career trajectories at three different levels. By analyzing these trajectories at

the city level, we have observed a skewed distribution in city attractiveness from scientists' point of view. Additionally, by taking a network perspective, we have found heterogeneous patterns for scientists' movements between less and more attractive cities. While by analyzing career trajectories at the affiliation and country level, we have demonstrated, respectively, the presence and absence of systemic temporal correlations. To conclude, with our analysis we have obtained a solid understanding of the mobility patterns of scientists.

Chapter 9

Scientists' mobility: A data-driven model

Summary

Building on the empirical findings of the previous chapter, we now propose an agent-based model that reproduces these findings on both the scientist and the network level. The model considers that agents have a fitness attribute and considers potential new locations if they allow increasing this fitness. Locations, on the other hand, rank agents against their fitness and consider them only if they still have a capacity for them. This leads to a matching problem which is solved algorithmically. Using empirical data to calibrate our model and to determine its initial conditions, we are able to validate the model against the measured distributions. By validating our model, we interpret the model assumptions as micro-based decision rules that explain the observed mobility patterns of scientists.[1]

[1]Based on [231]

9.1 Introduction

High-skill labor mobility is a crucial economic and political issue of our time. Modern economies rely on high skill labor to keep their competitive advantage [11, 19, 20]. For this reason, attracting and retaining scientists is becoming a central concern for migration policy [29]. In this work, we investigate the mobility of scientists by studying several forces that arguably drive their relocation choice. We propose an agent-based model that we calibrate and validate against real data. With this data-driven approach, we test if a set of minimal decision rules can explain observed mobility patterns of scientists.

Scientists are highly mobile individuals, a fact that has been true in the past and is becoming ever more important [71]. There is an expanding literature on the mobility of scientists. Many works have been focusing on the relationship between movements and scientific impact [55, 62, 196]. Other works analyzed scientist mobility across countries to determine the effects of policy [44] and to investigate aspects of the brain circulation phenomenon [3, 21, 195, 238].

Most works address scientist mobility at an aggregated level, i.e., they focus on bilateral flows between countries. At the same time, the need to understand the basic forces at scientist level underlying academic mobility has been highlighted by Appelt *et al.* [8], Fortunato *et al.* [59]. This need has been approached both empirically [63, 72, 239] and theoretically Mahroum [128]. Empirical works are traditionally based on survey data that provide only a small coverage of the global mobility of scientists and usually aggregated at the country level; while theoretical works are rarely validated against data.

In order to go beyond speculation in what drives the global academic mobility, we have reconstructed the global mobility network of scientists in Chap. 8. For doing this, we used the approach of Verginer and Riccaboni [238] that allows extracting geographical career trajectories of scientists using bibliographic data. In particular, we used the MEDLINE database, the largest open-access bibliographic database in the life sciences (see Sect. 8.2). Now, after reconstructing the mobility network, we propose an agent-based model to reproduce this network and other scientist-level properties. The model is explained in Sect. 9.2, while its calibration and validation procedure are respectively pre-

sented in Sect. 9.3 and Sect. 9.4, and they follow the data-driven approach of Tomasello *et al.* [224, 227], Vaccario *et al.* [233]. Finally, in Sect. 9.5, we further discuss the results from our simulations, analyze the limitation of the model, and provide some outlooks.

9.2 Overview of the agent-based model

In this section, we propose and define a model, which can reproduce the characteristic properties of the empirical mobility network discussed in Chap. 8. Precisely, we want to reproduce features both at *scientist* and *network* level. These are, on the scientists' level, (1) the distribution of move distances, Fig. 8.1(a) and (2) the "age at move" distributions, Fig. 8.1(b). Moreover, at the network level, we want to reproduce we want to reproduce (3) the distributions of the topological features shown in Fig. 8.4, i.e., local clustering coefficients, path lengths, degrees and degrees of neighbors.

We note that this is quite an ambitious goal since our model needs to correctly reproduce several *very different* system dimensions (i.e., scientists (micro), intercity (macro)). If the model is able to reproduce the described distributions, we have a strong indication that the interaction rules governing scientists and city interactions, capture a relevant aspect of the real mobility of scientists. The information available to the model during fitting does not imply the more complex validation measures. If we find that the simulated results agree with the empirical validation metrics, it means that the *interaction rules* are the reason for the observed patterns and good validation results.

We decide to develop an *agent-based model* because we want to model the mobility of scientists, as opposed to a system dynamics model in which we would merely reproduce the flows between different cities, on the aggregated level. This implies that *macroscopic features* describing the system or network level, such as the topological properties already discussed, have to be *emergent properties* arising from the agent dynamics.

9.2.1 Agents and Locations

Our model is composed of two entities, `agents` and `locations`. `Agents` represent scientists. Each agent i is characterized by three properties that change over time: its position, $r_i(t)$, its fitness, $f_i(t)$, and its years of activity $y_i(t)$. Time is measured in discrete simulation steps, each representing one year. When we start our simulations at time $t = 0$, which is chosen as the year 2000 below, we cannot assume that all agents also start to become active only then. Instead, agents have already been publishing before, which is included in $y_i(t)$. An agent that published its first paper in 1995 will have a $y_i(2000) = 5$ in this case. This becomes of importance when measuring the fitness of agents, $f_i(t = 0)$, as determined below.

`Locations` represent cities and host agents. In agreement with the dataset, we have $M = 5,485$ different locations. Each location K is characterized by three properties that can also partly change over time: its position R_K defined in real geographical space by means of longitude and latitude, its fitness, $F_K(t)$ (see Sect. 8.2.2), and the number of agents it hosts, $N_K(t)$ (see Sect. 8.2.2). Note that we take R_K and $N_k(t)$ from the available empirical data.For the fitness of a location, however, we do not take accumulated ranking values of institutions into account. Instead, we choose a different proxy for fitness, which is more consistent with our model: the fitness $F_K(t)$ is equal to the average fitness of all agents hosted in location K. This relates the problem back to defining the fitness of agents. But at the same time, it is in line with the ranking of academic institutions, which in essence is also determined by the fitness, or quality, of the scientists working there. In our model, we assume that the $F_K(t)$ are public information, just as the rankings are.

For the position $r_i(t)$ of an agent, we assume that at each time step the agent can be found in one of the available locations. So $r_i(t) = R_K$ where K is location of agent i is based at time t.

9.2.2 Model dynamics

Movement preferences. Our main modeling assumption is that agents prefer to work in locations that provide a higher fitness than the one where

they are currently based. These locations, however, can be very distant from the current place, which incurs higher relocation costs. Therefore, agents do not only take the fitness $F_K(t)$ of locations into account, but also the geodesic distance $\Delta_{i,K}(t)$ between the current location of i and any other location K. They combine this information in a *re-scaled fitness score* $\tilde{F}_{i,K}(t) = F_K(t)/(\Delta_{i,K}(t))^b$ for each location K. b is a model parameter, used to weigh the impact of spatial distances. The bigger b, the more important any spatial distance becomes.

By ranking the values $\tilde{F}_{i,K}(t)$ from high to low, each agent then obtains an *individual* ranking that reflects its preferences where to move next. Agents in L will consider only those locations where $F_K(t) > F_L(t)$, i.e., where the average fitness of scientists is larger than the average fitness of scientists in their city. Agents' preferences can be summarized in a ranking vector $\vec{V_i}$ where its K-component, $(\vec{V_i})_K$, represents the preference for location-K and it can be written as

$$(\vec{V_i})_K = \Theta\left(F_K(t) - F_L(t)\right) \frac{F_K(t)}{(\Delta_{i,K}(t))^b} \tag{9.1}$$

Movement decisions. Agents only come up with a ranked list of possible locations they would consider to move to (and we can assume that they send applications to the academic institutions in these locations). However, agents do not decide where to move. This decision, whether or not to accept the agent, is taken at the location.

A location K will accept new agents only if its capacity allows so, which is defined by $N_K(t)$, the number of scientists empirically observed at a given location. External factors, such as the growth of academic institutions, are implicitly considered in the observed change of $N_K(t)$. As we found out, the $N_K(t)$ are rather stable over time. This implies that, after some transient periods in our simulations, locations have the capacity to accept incoming agents only if agents at K have been accepted somewhere else and move there.

Because, dependent on the individual ranking of agents, some locations obtain more applications than the capacity allows them to accept, each location ranks the qualified agents according to their fitness $f_i(t)$. Available slots are filled starting from agents with higher fitness values until the capacity

$N_K(t)$ is reached. Precisely, if $f_i(t) > F_K(t)$, location K considers agent i with probability $p = 1$ because this allows location K to increase its fitness $F_K(t)$. If $f_i(t) \leq F_K(t)$, location K considers agent i only with a probability $p = (f_i(t)/F_K)^s$ where s is our second model parameter. Note that for a high value of s, locations become more selective. We express the probability of a location-K to consider an agent-i as a piecewise function of the fitness of the agent and of the location:

$$p(K,i) = \begin{cases} 1 & f_i(t) > F_K(t) \\ (f_i(t)/F_K(t))^s & \text{otherwise} \end{cases} \tag{9.2}$$

For $s \to \infty$, $p(K,i) \to \theta(f_i(t) - F_K(t))$, i.e. a location considers an agent if and only if its fitness is higher compared to the average fitness of the agents currently hosted in the location. While for $s \to \inf$, $p(K,i) \to 1$, i.e. a location considers all agents independently of their fitness. In Fig. 9.1, we visualize the basic rules of our model.

Matching problem. In our model agents rank locations, while locations rank agents. To match locations and agents, we have to solve a matching problem similar to the stable marriage problem. However, our problem is slightly different as a location can accept more than one agent until the capacity $N_K(t)$ is reached. To solve this matching problem, we use the established NRMP-algorithm developed by the National Resident Matching Program (NRMP) for matching medical students to U.S. training programs. After the matching is completed, only the agents that have been matched to a location will move. If an agent i has moved to a new location K, we update its position vector, $r_i(t+1) = R_K$, and keep its fitness constant, $f_i(t+1) = f_i(t)$.

Fitness dynamics. We have to decide what happens to all those agents that are not accepted at a new location. Here, we consider that the agent stays at its current location, i.e. $r_i(t+1) = r_i(t)$, and uses the time step to further improve its fitness, $f_i(t)$. For this, we assume a stochastic dynamics, precisely an additive stochastic process with a variance proportional to the fitness of the current location. This implies that it is not guaranteed that

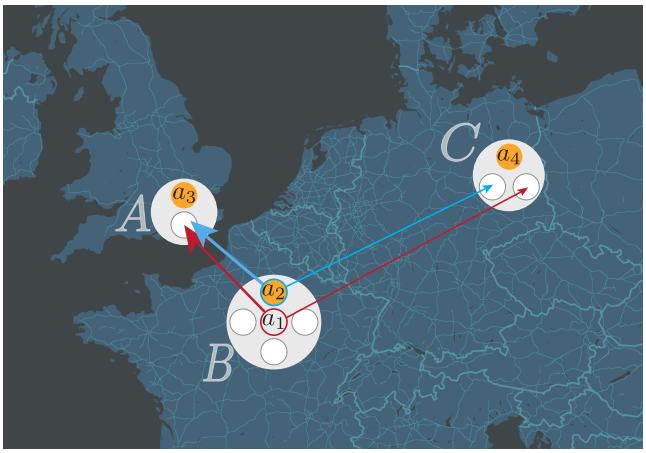


Figure 9.1: The agents (a_1, a_2, a_3 and a_4) are all hosted in three locations, A, B or C, that represent respectively London, Paris and Berlin. Each location has a maximum number of available positions illustrated by some small slots, $N_A = 2$, $N_B = 4$ and $N_C = 3$. In this image, agents a_1 and a_2 compute the rescaled fitness of the available locations (A and C) and rank these locations accordingly. Here, we have assumed that A and C have the same fitness ($F_A(t) = F_C(t)$), but A is closer to B than C is ($\Delta_{i,A} < \Delta_{i,C}$ for $i = 1, 2$). For this reason both a_1 and a_2 express a preference for A over C. Since location A has $N_A = 2$ and one position is already taken, A must decide to host either a_1 or a_2. Location A will decide depending on the fitness of a_1 and a_2.

agents will increase their fitness when not moving. At the end of each time step, we update the fitness of locations, $F_K(t)$, by averaging over the fitness $f_i(t)$ of all those agents that are currently based there.

9.2.3 A data-driven model

We use the empirical data not only as an *input* to our model, but also for *calibrating* and *validating* it. As *input*, we take six observed quantities: three at the city level and three at the scientist level, to determine the initial conditions of our model. As the starting year $t = 0$, we take 2000.

From each city we take its geographical position and the number of scientists in year 2000. We assign these quantities to locations to characterize their R_K and $N_K(t=0)$. The initial fitness value of a location, $F_K(t=0)$, is determined by averaging over the fitness values of those agents based in the given city in 2000.

From each scientist, we take its geographical position (in a given city), its academic impact and the years of activity already passed until 2000. We assign these quantities to agents to characterize their $r_i(t=0)$, $f_i(t=0)$ and $y_i(t=0)$. The *academic impact* is proxied by the papers that a scientist has co-authored during the last two years of activity. Precisely, we assign to each paper a score equal to the impact factor of the journal[2] where it was published divided by the number of co-authors. Then, for each scientist, we sum the scores of the papers he/she has co-authored between 1998 and 2000. This defines the starting fitness of agents, i.e. $f_i(t=0)$.

We then run the agent-based model using parallel updates of all agents per time step, taking as evolving quantities only the values of $N_K(t)$ into account. To do so, we still have to determine the two free parameters of our model, b, which weighs the impact of spatial distances for the individual rankings of agents, and s, which weighs the flexibility of locations to still accept agents with a fitness less than the fitness of the location. Determining b and s is done during the model calibration.

9.3 The calibration procedure

To *calibrate* the model, we use two empirical distributions: the inflow and the outflow distributions shown in Fig. 9.2(a,b). Note that for this manuscript, we calibrate our model considering only cities and scientists present in three countries: France, Germany, and the United Kingdom. To determine the model parameters b, s from that, we use an established approach in agent-based modeling [233], machine learning [22, 116, 182], and computer simulations in

[2]The Journal impact scores are taken from Scimago, see 2 for details.

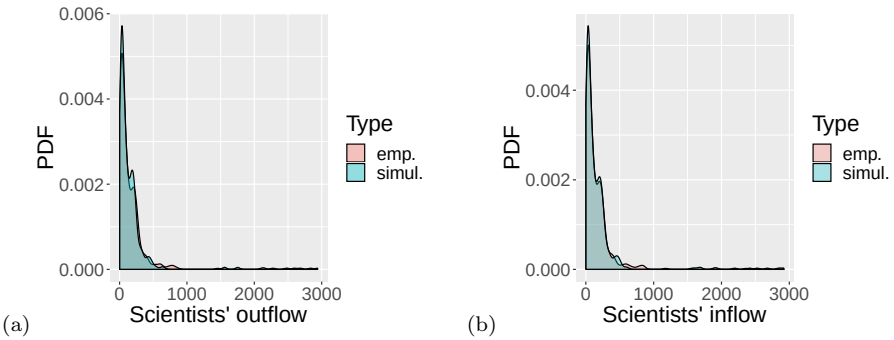


Figure 9.2: Distributions of (a) inflow of scientists into any city, (b) outflow of scientists from any city. (red) indicates the empirical distributions, (blue) the (optimally) simulated distributions obtained from the calibration of our agent-based model.

general [114]. It combines two elements: (a) a grid search and (b) a performance score.

The grid search consist of an exploration of the (low dimensional) parameter space through computer simulations. For b the values $\{0.005, 0.01, 0.05, 0.1, 0.5, 1.0, 5.0\}$ are considered, for s the values $\{0.05, 0.1, 0.5, 1.0, 5.0, 1.0, 10.0\}$. Recall that for $b \to 0$, distances do not play a role when computing the re-scaled fitness scores $\tilde{F}_{i,K}(t)$. Already for $b = 0.005$, we obtain re-scaled scores with differences smaller than 5% even if the distances have differences of three orders of magnitude. While for $b \to \infty$, distances are the dominant factors in $\tilde{F}_{i,K}(t)$, but already for $b = 5$ this is the case: Given an agent i in a location L and other two locations K and K' such that $F_K(t) = 100F_{K'}(t)$ and $\Delta_{i,K} = 3\Delta_{i,K'}$, then $\tilde{F}_{i,K}(t) < \tilde{F}_{i,K'}(t)$ for $b = 5$. This means that the city with 100 lower fitness (but one third) closer is preferred already when $b = 5$. Also note that for $s = 0.05$, an agent with almost fitness close to 0.01 would still be accepted with a probability of almost 80% by a location with average fitness equal to 1. While for $s = 10.0$, an agent with fitness close to 0.8 would be accepted with a probability smaller than 10% by a location with average fitness equal to 1. Thus, we can assume that the chosen parameter space encloses almost the full space of events that we can model.

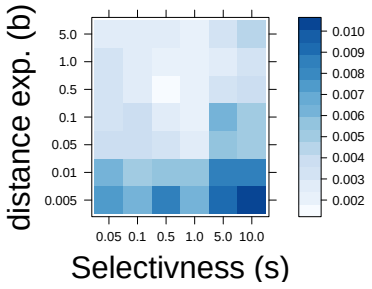


Figure 9.3: The heat-map shows the results of the grid-search on the two parameters s and b. The color of each cell corresponds to a p for a given (b, s) pair as described in Eq. (9.3). The optimal parameter pair (b^{opt}, s^{opt}) is $(0.5, 0.5)$.

For each parameter combination, we simulate our agent-based model and obtain two distributions for the inflow and outflow, as shown in Fig. 9.2(a,b). We now have to determine the optimal combination of (b, s) that matches the empirical distributions best. For this, we use a performance score based on the Kolmogorov-Smirnov(KS) statistic [108]. Precisely, for each combination of parameters (b, s), we compute the KS-statistic between the empirical and simulated distributions of inflow, $D_1(b, s)$, and of outflow $D_2(b, s)$. We then define the performance score as $1/(D_1(b, s) \times D_2(b, s))$, such that the optimal combination (b^{opt}, s^{opt}) maximizes this score. We can write this with the following equation:

$$p^{opt} = (b^{opt}, s^{opt}) = \left(\arg\max_{b,s} \frac{1}{N} \sum_{k}^{N} D_1(b, s) \times D_2(b, s) \right)^{-1} \tag{9.3}$$

where N is the number of simulations, $D_1(b, s)$ is the Kolmogorov-Smirnov statistic between the empirical and simulated distributions of city outflow. $D_2(b, s)$ is the Kolmogorov-Smirnov statistic between the empirical and simulated distributions of city inflow.

In Fig. 9.3, we report a heat-map showing the exploration of the parameter space. For each combination of parameters, 10 simulations are run. We find as *optimal parameters* $p^{opt} = (s^{opt}, b^{opt}) = (0.5, 0.5)$. This means that both

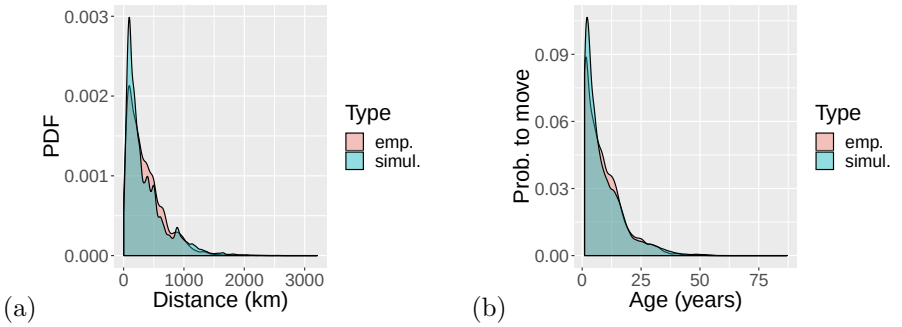


Figure 9.4: (a) Distribution of movement distances of scientists. (b) Distribution of movements dependent on the (academic) age of scientists. (red) indicates the empirical distributions, (blue) the distributions that are obtained from our agent-based simulations. The distributions are obtained from the frequencies using the default smoothing of `ggplot2` in `r`.

selectiveness and distances better reproduce the empirical data when they give a sub-linear contribution. The comparison between the empirical and the simulated distributions is shown in Fig. 9.2 (a,b). The close match demonstrates that our model is correctly calibrated. Some smaller differences are discussed in Sect. 9.5.

9.4 Results of the agent-based simulations

The calibrated agent-based model has to prove its evidence in that it is able to reproduce also the whole set of empirical findings that have *not* been used during the calibration procedure. If that is the case, the model has been *validated*. As already mentioned, we will verify this for two distributions on the level of scientists and four distributions on the level of the movement network.

The results of the validation are shown in Fig. 9.4 and 9.5. To allow for a direct comparison, we plot the empirical data in red and the simulations in blue. We can report a very good match of all distributions both on the level of scientists and on the network level. Specifically, on the scientists' level, we

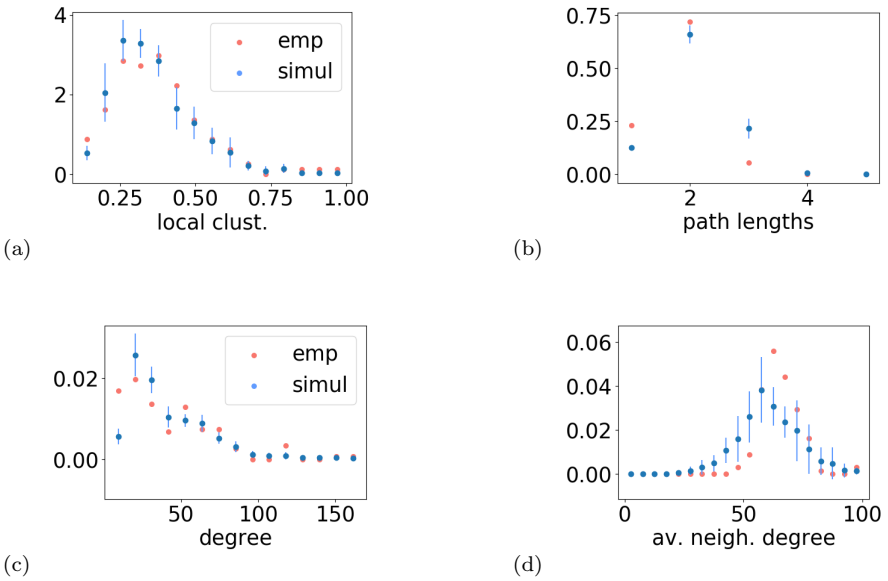


Figure 9.5: Distributions of (a) local clustering coefficients, (b) path lengths, (c) degrees, and (d) degrees of neighbors. (red) indicates the empirical distribution, (blue) the distributions that are obtained from our agent-based simulations. The error bars correspond to the standard deviations of the measures computed on 10 different realizations of the simulated mobility network.

are able to reproduce the two distributions of *movement distances* and of *age* when moving, see Fig. 9.4(a,b).

On the network level, we are able to reproduce the four distributions of *clustering coefficients*, *path lengths*, *degree* and *neighbor degree*, see Fig. 9.5(a,b,c,d). We emphasize that these results are far from being trivial. As we start with an agent-based perspective, the results of our simulations refer to *career trajectories* of individual agents. From these, we have to reconstruct an aggregated mobility network as described in Sect. 8.2.2. Our simulation results for the network topology are reported for these simulated networks.

In conclusion, we report that our agent-based model captures the different features of the empirical data very well, both on the scientists' and the network level, without using direct information from these for the calibration.

9.5 Discussion and outlook

The important contribution of this chapter is an agent-based model that allows us to reproduce the empirical findings discussed in Chap. 8, both on the level of scientists and the level of cities. In our model, we assume as most relevant factors geographical distances, academic importance, and selectiveness of cities. This model uses as input only variables that can be proxied by the available data. This extends in particular to the notion of academic importance, denoted as "fitness", assigned to agents, which is proxied initially from the available publications. The "fitness" of locations, another ingredient of the model, can be then obtained by averaging over the fitness of agents at the particular locations.

The agent-based model further uses only straightforward assumptions as rules to determine the movement of agents. Agents rank all locations according to their fitness and their distance to the current location. However, they do not decide about the movement. The decision is taken by the locations using information about the fitness of the agents and capacity constraints for the hiring of new agents. In essence, this poses a matching problem and can be related to similar problems discussed in the literature.

Our agent-based model only considers two free parameters, which need to be calibrated against the available data: b weighs the spatial distance between the current location of an agent and any other location, s weighs the selectiveness of locations when accepting agents that have a fitness below the average obtained for that location. We find as *optimal parameters* $(s^{opt}, b^{opt}) = (0.5, 0.5)$. This means that both selectiveness and distances better reproduce the empirical data when they give a sub-linear contribution.

Using the model calibrated with the optimal parameters, we are able to reproduce the available empirical data very well. At the same time, there are minor differences between the simulated and the empirical distributions that still needs to be quantified. In a nutshell, they are also due to the fact that the simulations are only done with 22,100 agents, while the data are obtained from 3.5 million scientists. These discrepancies become noticeable if we plot the network of scientists' mobility on the European scale, only, as it is shown

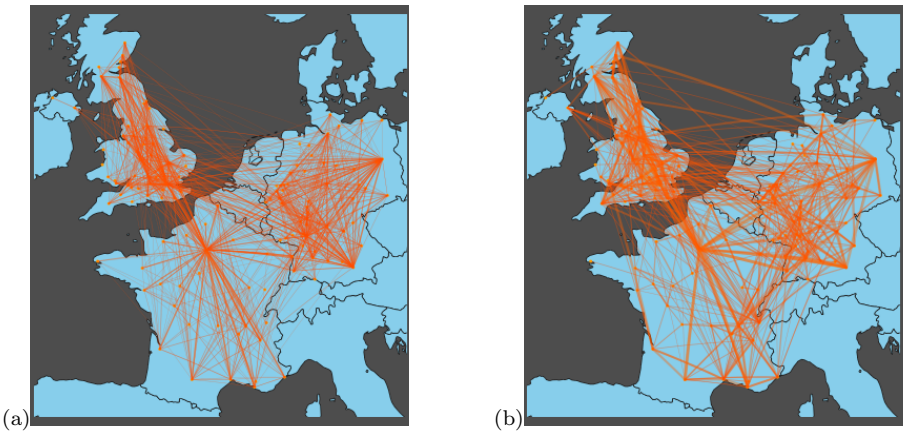


Figure 9.6: Empirical and simulated mobility networks for France, Germany and UK. The empirical network (a) depicts the flows between cities, the thickness of links indicates their magnitude. The map in (b) depicts one realization of the ABM with optimal parameters.

in Fig. 9.6. We observe that the empirical network in Fig. 9.6 (a) shows more pronounced hubs than the simulated network shown in Fig. 9.6 (b). Specifically, in the empirical network, significantly more French cities are linked to Paris compared to the simulated one.

Finally, we stress that more factors are influencing the relocation choices of scientists than explicitly covered in our model. For example, quality of life, better networking opportunities, or higher salaries might be relevant factors here. The more remarkable is the fact that our model, even at this level of detail, works considerably well. Going forward, we want to understand how other factors might explain the empirical core-periphery structure. Accounting for this in the simulation might help to recreate more subtle migration patterns (i.e., a very central Paris). Moreover, we could replace the geographical distance between locations by travel time between cities, since this is most likely how humans estimate travel effort. Additionally, we find in Verginer and Riccaboni [238] that migration is marked by national borders and cultural/language similarity. These findings would be exciting features to reproduce further. Last, but not least, we have currently centered our analysis around

France, Germany, and the United Kingdom, but will spend further effort on the global migration network for future research.

When we started this thesis, we aimed at developing an ABM to reproduce scientists' career trajectories and also their temporal correlations (**RQ7**). However, we have found that such correlations are present at the affiliation level, but not at the country or city level (see Chap. 8). This fact has brought additional complexity to the empirical phenomenon of scientists' mobility and has partially invalidated the assumption behind **RQ7**. One could have ignored this fact and simply developed an ABM to reproduce the (second-order) temporal correlations at the affiliation level. Instead, we have performed an extensive analysis to identify the individual and global properties of scientists' career trajectories, and then, we have developed an ABM to reproduce them. Now this model provides a solid foundation to explore the additional microscopic rules that capture the presence of temporal correlations in scientists' career trajectories at the affiliation level and their absence at the country level.

In summary, with our research, we have provided the first agent-based model reproducing the mobility of scientists. In a data-driven approach, our model has been calibrated and validated against data, and we have found an extremely good match between simulations and empirics. With this, we show that minimal decision rules capture many complex features of the observed mobility of scientists. In addition, we have quantified the relative importance between geographical distances and academic attractiveness from the perspective of a scientist trying to relocate.

Chapter 10

Conclusions

The main purpose of this thesis was to study knowledge in socio-technical systems. In particular, we have dedicated our attention to two systems, Academia, and R&D alliances. We have found a strong interplay between knowledge and the actors of these systems (i.e., scientists and firms). On the one hand, knowledge is exchanged and created collaboratively by these actors, and hence, their interactions determine the structure and diffusion of knowledge. On the other hand, these interactions are also determined by the knowledge of the actors. Indeed, scientists, as well as firms, decide on their collaborators depending on the knowledge that these particular collaborators own.

To quantify the effect of social interactions on knowledge, we have studied the structure of explicit scientific knowledge, i.e., of knowledge encoded in scientific publications. By representing these publications and their citations as a network, researchers have developed citation-based indicators that rank publications according to their impact. We have developed a statistical test to see whether these rankings were "fair", meaning that we check whether these indicators favor or disfavor papers of a specific age or from specific scientific fields. Similar to previous works, we have found that none of the existing indicators were fair and with our new test, we quantitatively asses this.

We argue that there are two reasons why established indicators fail to capture the impact of publications across scientific fields and time. The first reason is the well-known cumulative effect of citations, i.e., that older papers have

more time to attract citations, and hence, they have an advantage compared to younger ones. Interestingly, we have found that such an effect is stronger for indicators based on eigenvectors centralities compared to indicators based on citation count. The second reason is that the distributions of paper citations from different scientific fields have not only a different first moment (i.e., the average number of citations) but also different second and higher moments. This indicates that these distributions have probably different functional forms and there is no evidence for claiming the existence of a universal distribution across scientific disciplines. We argue that the differences in the distributions reflect the presence of different citation norms in different scientific fields. In other words, social norms in scientific communities influence the structure of the citation network to the point that all the analyzed citation-based indicators are inadequate to compare paper impact across fields and time.

To further quantify the effects of social interactions on knowledge, we have analyzed the mutual exchange of knowledge among collaborating firms and scientists. To do this, we have considered patents and publications authored by these actors and assigned them positions in knowledge spaces. The knowledge spaces for firms and scientists were defined using real-world classification schemes for patents and publications, respectively. In these spaces, we have introduced a notion of distance capturing knowledge similarity: two actors have more similar knowledge when they are close in the knowledge space. By analyzing the distribution of changes in distances between actors, we have found that both scientists and firms approach each other after collaborating. Hence, we have found evidence for knowledge exchange.

After analyzing the effect that collaborations among firms and scientists have on their knowledge, we have analyzed the opposite effect, i.e., we have studied the role of their knowledge in determining collaborations. In both R&D alliances and scientific co-authorship activities, we have found that knowledge has a central role. This was captured mainly by two observations. *First*, in both systems the distribution of pre-knowledge distances cannot be expected at random. This indicates that firms and scientists collaborating decisions correlate to their differences in knowledge. In the economic domain, we have found that most alliances are established between firms with different knowl-

edge, but not too different. While scientists co-author publications mostly with other scientists with either very similar or different knowledge.

Second, we have used the distribution of pre-knowledge distances to reproduce collective properties depending on collaborations. By using this distribution as input for two ABMs, we have reproduced the knowledge exchange between collaborating firms and the productivity of scientists. Also, for the knowledge exchange, we have found that this is rather small, meaning that firms do not move a lot in the knowledge space after collaborating. Hence, knowledge is rather a determinant than a consequence of R&D collaborations. While by considering the heterogeneous publishing activity of scientists and the empirical distribution of pre-knowledge distances, we have reproduced scientists' productivity. Precisely, we have reproduced its non-linear trend as a function of the knowledge distance between collaborating scientists.

Last, but not least, we have provided a new perspective on how to study the structure of knowledge and its transfer. Building on advances in data mining and network analysis, we have used a path abstraction to capture temporal correlations in citation data and in scientists' career trajectories. We have shown that these correlations (i) totally change the ranking of journals computed using PageRank, and (ii) allow to better capture similarities among journals. Additionally, we have found that temporal correlations are statistically significant when studying scientists' careers at the affiliation level, but not at the city or country level. This indicates that knowledge transfer from a geographical point of view can be represented using a traditional network perspective only at the country and city level. While at the affiliation level, this should not be done. Note that nowadays the decision to hire a scientist or to fund research projects of scientists often depends on the journals where scientists publish their works and on their mobility. Hence, our analyses based on a path abstraction provide relevant insights for decision-makers developing policies for public and private research.

10.1 Scientific contributions

For a long time, understanding the structure and evolution of knowledge has been a central focus only for few disciplines, like philosophy. Nowadays, with the availability of large data sets on how knowledge is and has been produced, we need a multidisciplinary effort to address such a topic. On one side, there is a need for new methods to filter and process this data. This aspect is one of the main focuses of scholars working in computer science and information science. One the other side, there is a need for new models to interpret the filtered and processed data. This second aspect is usually tackled by scholars from sociology, network science, management, and econophysics. With our study of knowledge in socio-technical systems, we have developed new methods and models that contribute to information science, network science, and agent-based modeling.

10.1.1 Contributions to Information Science

Our primary contribution to information science is the statistical test developed to quantify biases of rankings. Given an indicator that ranks items, our test checks whether this indicator is favoring or disfavoring items belonging to specific categories. Note that our test can simultaneously quantify multiple biases, and given its analytic formulations, we can also compute the contribution of each bias to the total one. Furthermore, our test is general as it can be extended to check for many types of biases. For example, we have used it to check whether our multidisciplinary indicator was favoring journals belonging to specific disciplines. In addition to our test, we have developed two new indicators of paper impact that are less biased compared to all the other analyzed indicators. Then, we have defined a new procedure to project the empirical flow of knowledge from the paper level to the journal level. With this procedure, we have computed not only new rankings for journal impact but also created new similarities measures. Note that our methods to quantify biases, impact, and similarity not only have further application in scientometrics (e.g., in patent citation analysis), but can also be used to improve information retrieval algorithms.

10.1.2 Contributions to Network Science

The network perspective is at the base of the representation and analysis of many systems. Building on recent advances in data mining and network science, we have shown how this perspective fails to represent citations at the journal level and scientists' career trajectories. Indeed, when using a network perspective to analyze citations, we aggregate them and introduce a fictitious knowledge mixing among papers belonging to the same journal. While in scientists' career trajectories, we discard statistically significant temporal correlations in the sequences of institutions where scientists have been working. To detect and overcome these problems, we have used a newly introduced path abstraction and higher-order network models. With all this, we contribute to the understanding of the limitations of network models and how to overcome them by using higher-order network models.

10.1.3 Contributions to Agent-Based Modelling

We have developed two agent-based models capturing the effect of knowledge on collaborations in scientific and R&D collaborations. Note that the microscopic rules of the developed models have been chosen depending on previous empirical findings and theoretical works ranging from psychology, sociology, and economics. Additionally, differently from most existing ABMs, our models are data-driven, meaning that we have used data as input as well as for calibrating and validating them. By this, we have developed models that are not just simulations reproducing collaborations, but rather models that allow us to understand and explain them. Finally, we have created a data-driven ABM that correctly reproduces the mobility of scientists at the city level. This model has two parameters determining the influence of distance and selectiveness of cities in the relocation process of scientists. With only these two parameters, we have captured the collective movements of scientists that transfer between cities in Germany, France, and the UK in order to work.

10.2 Outlooks

To define possible outlooks for this Ph.D. work, we ask the following two questions: *Given the results obtained, what are the next steps to either generalize or disprove them?* And *given the difficulties and problems we have faced, what could have we done differently?* The former question has a clear relation to the possible outlooks of a project, while the latter has a more subtle relation. By defining the faced difficulties, one can identify the shortcomings of the used procedure and developed methods. By this, one can propose new procedures and methods that could overcome the identified shortcomings. In the following, we address the above posed questions by providing new possible applications of the developed methods and some methodological outlooks.

Descriptive and Generative models. We have concentrated on generative models in this thesis. Generative models are quite restrictive and usually computationally expensive, and hence, they require a parsimonious mindset when developing them. Indeed, we have excluded the role of many factors in our analysis based on intuition or results from previous research. This was time-consuming and could have also influenced us to discard relevant factors. Instead of doing this, we could have used descriptive models (e.g., models obtained from linear regressions or random forest classifiers), at the begging of our analyses to identify the most relevant descriptive variables. Then, we could have used the identified variables to develop our generative models. With these, we could have identified the most relevant descriptive variables and then, could have used them to develop our generative models. In other words, we could have supported the development of generative models by introducing a feature selection process based on descriptive models. By applying this approach, we could confirm or question the choices that we have done, for example, our decision to use pre-knowledge distances as input for our ABMs. Also, descriptive models could suggest which other features of the data to add to improve our models.

Testing for fairness. Given the increasing use of data in governing science [94] and society in general (e.g., see the Social Credit System in China),

quantitative indicators are more and more used in real-world evaluation processes. These indicators are often used as they are considered objective and can avoid bias tendency of evaluators during these processes. However, we have shown with our statistical test that quantitative indicators (of paper impact) also have biases. Hence, we have to consider the possibility that our test will be applied to identify less biased indicators to be then used in evaluation processes. To avoid the miss-interpretation of our test and its missus in such crucial processes, we stress a couple of points. First, our test was developed to quantify the different biases of rankings and to show that citation norms vary in time and across scientific communities, and not to support evaluation processes. Its application to this different setting still needs to be investigated, and it should be done carefully. Second, when defining the bias of rankings, we had to define a null expectation, i.e., how we *believe a fair* ranking should be. This null expectation is introducing a subjective component that poses a limitation of our test and its possible missus. Hence, we strongly suggest that the null expectation should be clearly stated and justified by any researcher or evaluator using our test. Additionally, after an indicator is considered to be fair by our test (or any other), it should not be applied blindly during an evaluation process. With these two suggestions, we pose our self in line with [94] where the authors state 10 scientometrics principles of research and evaluation processes (especially see points 1) and 4)).

Multi-order Graphical models and Infomap. Even though these two methods are deeply different in their usage, they also share many similarities. `Infomap` is a clustering algorithm on networks, while multi-order graphical models represent sequential data using a mixture of higher-order network models. At the same time, both methods are based on the idea of finding a parsimonious representation of flows on networks, i.e., they try to compress sequential data. With these compression processes, we can define an optimal order that is the maximum (Non-Markov) memory statistically significant in the analyzed sequential data. Understanding for which data generating process `Infomap` and multi-order graphical models detect the same order is still an open problem. This problem can also be stated with the following question: Given that `Infomap` is detecting communities, when are these communities so

statistically significant that both `Infomap` and multi-order graphical models detect the same order?

Modelling citation and publishing norms. In this thesis, we quantitatively showed that the rankings produced by different indicators either favor or disfavor specific scientific fields. This provides evidence that publishing and citation norms vary across communities. At the same time, we have not explained how these norms are different. We find particularly interesting to study this by first further analyzing differences across communities and then, to model these using different citation norms. For example, we could check whether more applied disciplines cite more recent publications compared to less applied ones. Then, we could use this phenomenon (if present) to model a stronger bias favoring older papers in less applied disciplines.

Knowledge management in companies. Our methods allow us to quantify the interplay between knowledge and interactions among actors in socio-technical systems. Given this interplay, we envision the extension of these methods to support the knowledge management in real companies. For example, consider software developer companies and assume that we define a knowledge space depending on software codes. Then, different developers in a company can have different positions depending on the code they write, but they also interact when working on the same project. By monitoring developers' positions in the knowledge space and by using these positions to organize teams correctly, managers could better administrate and govern their companies. For example, they could avoid the formation of knowledge island or of teams formed by developers with too overlapping knowledge[1]. The extension of our methods to this application needs for new data that it is not easily available, i.e., a tractable representation of co-edited codes that allows us to assign knowledge positions to developers. At the same time, such data is becoming increasingly available thanks to the development of new data mining tools, such as the one of Gote *et al.* [78].

[1]Note that this is the central research topic of Cristoph Gote, a Ph.D. candidate in the Chair of Systems Design.

Scientists mobility. As one objective of this thesis, we had to reproduce scientists' career trajectories with an ABM. We have done this by reproducing their movements at the city level, and we have tested our model only against data of scientists working in the UK, Germany, and France. However, we still have not reproduced their movements at affiliation level where we have found the presence of temporal correlations in their trajectories. In order to do this, we first want to understand what minimal rules create these temporal correlations when changing the aggregation level of the data. Then, we have to map these rules to the real-world case of scientists' movements in order to model their careers at different levels. With this model, we would provide a new unique tool for policymakers to analyze the brain circulation phenomenon at the affiliation, city, and country level.

Appendix A

KDD Cup data

We do not distinguish the publications by their type (paper, review, book, etc.). Further, we also do not differentiate between different types of journals and take into account all of them: for example, we do not distinguish between a citation coming (or going) from (or to) a letter or a book. We argue that it is important to keep various types of journals and publications because different fields adopt not only different citation norms, but also different ways to communicate their results. For example, computer science researchers commonly publish results in conference proceedings, while physics authors tend to prefer articles or letters. At the same time, we are aware that different types of publications might have different citation characteristics. However, good indicators should ideally be able to account for heterogeneity among publications and citation norms across different communities and produce unbiased rankings without the need for arbitrary choices about which types of articles to include in the analysis. In addition, similarly as [249] and differently from [179], we do not exclude publications which do not receive citations.

The `MAG` has a field classification scheme with 4 hierarchical levels. The field assignment is based on an internal algorithm that uses a machine learning approach [213]. In our work, similarly to [179], we are only interested in impact metric normalization at the most coarse-grained level. To this aim, in our analysis, we focus only on the 19 main fields as listed in Table A.1. Discussing the possible limitations of the classification approach by `MAG` and

Field	Publication count	Mean citation count
Art	233 251	3.90
Biology	5 847 554	9.67
Business	613 827	4.81
Chemistry	6 204 531	7.28
Computer Science	4 080 636	6.13
Economics	2 252 921	5.56
Engineering	3 011 763	5.10
Environmental Science	315 465	12.63
Geography	288 338	6.92
Geology	1 825 707	7.88
History	390 144	5.53
Materials Science	2 063 474	6.12
Mathematics	4 551 453	5.87
Medicine	5 061 990	7.90
Philosophy	787 649	5.05
Physics	6 976 644	5.55
Political Science	144 473	2.51
Psychology	2 861 813	8.23
Sociology	1 784 695	5.39
Total	49 296 327	–
Total (no multiple)	18 193 082	6.42

Table A.1: Main Fields The 19 main fields identified by Microsoft Academic Graph with their number of publications and average citations. The second to last row reports the total number of publications considering multiple times publications that belong to more than one fields, whereas the last row reports the total number of unique publications and the average citation count.

the dependence of our results on the adopted classification scheme is a relevant subject for future research.

We only included in the analysis publications for which the following information are available: (1) unique identifier (ID); (2) complete publication date (yyyy/mm/dd); (3) DOI or journal-id, in order to be able to retrieve the publication; (4) assignment to at least one of the main 19 fields. We discard from our analysis publications for which one or more of these four properties are missing.

With this filtering procedure, we obtain $N = 18\,193\,082$ unique publications and $E = 109\,719\,182$ citations.

Appendix B

First moment rescaling

According to [179], the distribution of the c^f indicator is log-normal:

$$F(c^f)dc^f = \frac{1}{\sigma c^f \sqrt{2\pi}} e^{-[\log(c^f)-\mu]^2/2\sigma^2} dc^f \tag{B.1}$$

where $\mu = -\sigma^2/2$ and σ is fitted from the data. When Eq. (B.1) is verified, then also the distributions of citation count, c_i, for all the individual fields, i, are lognormal:

$$F(c_i)dc_i = \frac{1}{\sigma c_i \sqrt{2\pi}} e^{-[\log(c_i)-\log(c_0)-\mu]^2/2\sigma^2} dc_i \tag{B.2}$$

where c_0 is the mean of c_i. For lognormal distributions the variance is proportional to the square of the mean and the constant of proportionality is $(e^{\sigma^2} - 1)$. From Eq. (B.2), we see that the citation counts c_i are distributed lognormally with mean $e^{\mu+\log(c_0)+\sigma^2/2}$ and variance $(e^{\sigma^2} - 1)e^{2\mu+2log(c_0)+\sigma^2}$. Recalling that $\mu = -\sigma^2/2$, we have that the mean is c_0, as it is expected, while the variance becomes $(e^{\sigma^2} - 1)c_0^2$. Thus, when Eq. (B.1) is verified, the variance of the empirical distribution of the citations for each field has to be proportional to the square of the mean citation count. Moreover, the constant of proportionality has to be $(e^{\sigma^2} - 1)$ for every field and year.

The analytic result just presented is in line with the Eq. (C.1) given in Appendix C of [130]. There it is shown that a rescaling procedure based on diving the original score by their first moment works if the ratio between standard

deviation and mean is constant. In the case of the relative citation ratio, we can calculate analytically such constant using the lognormal distribution and obtain the fitting parameter σ^2.

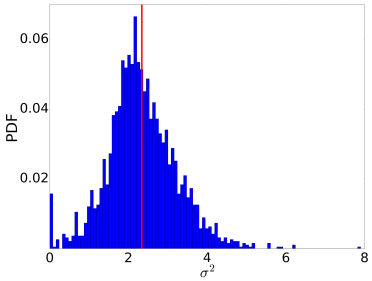


Figure B.1: Distribution of σ^2 obtained by calculating the empirical ratio, $r = (e^{\sigma^2} - 1)$, between the variance and the square of the mean citation count of each field and year.

In Fig. B.1, we report the distribution of σ^2 obtained by calculating the empirical ratio between the variance and the square of the mean, r, and by inverting the relation $r = (e^{\sigma^2} - 1)$ for every field and year. If the universality claim was correct, we would expect a narrow distribution of σ^2. By contrast, we find that σ^2 ranges between 0 and 8 across different fields and years. We argue that the broad range of σ^2 is the reason why the first moment rescaling introduced in [179] does not work in the analyzed dataset.

Appendix C

The Mahalanobis distance

The Mahalanobis distance ($d_{\mathcal{M}}$) is an established measure in statistics which generalizes the concept of z-score to multivariate distributions by taking into account also possible correlations between the random variables [127]. Its definition reads

$$d_{\mathcal{M}}(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T \mathbf{S}^{-1}(\vec{x}-\vec{y})} \tag{C.1}$$

where $\mathbf{S}^{-1}$ is the inverse of the covariance matrix, $\vec{x}$ and $\vec{y}$ are two vectors containing the random variables. When the covariance matrix is diagonal, i.e., the random variables are not correlated, than the $d_{\mathcal{M}}$ is equivalent to the square root of the sum of the squares of the z-scores.

In Section 3.5.2, we have used Eq. (3.5), an expression for the $d_{\mathcal{M}}$ valid when the covariance matrix comes from a Multivariate Hypergeometric Distribution (MHD), i.e., when the elements of the matrix are

$$S_{ij} = \left(\delta_{ij}(K_i(N-K_i)) - (1-\delta_{ij})\, K_i K_j\right)\gamma \quad \forall i,j = 1,\ldots,F-1 \tag{C.2}$$

where δ_{ij} is the Kronecker delta, K_i is the number of papers of category i, $N = \sum_i^F K_i$ is the total number of papers, F is the number of paper categories, $\gamma = \frac{n(N-n)}{N^2(N-1)}$ and n is the number of sampled papers. It is worthy to remember that even though we have F different categories, we only have $F-1$ degrees of

freedom. Here, we derive Eq. (3.5) for the case of a MHD in three dimensions, i.e., for $F = 3$. In this case, the covariance matrix is 2×2:

$$\mathbf{S} = \gamma \begin{pmatrix} K_1(N - K_1) & -K_1K_2 \\ -K_1K_2 & K_2(N - K_2) \end{pmatrix} \tag{C.3}$$

and the inverse of the covariance matrix is

$$\mathbf{S}^{-1} = \frac{1}{\gamma det(\mathbf{S})} \begin{pmatrix} K_2(N - K_2) & K_1K_2 \\ K_1K_2 & K_1(N - K_1) \end{pmatrix} \tag{C.4}$$

where $det(\mathbf{S}) = K_1(N - K_1)K_2(N - K_2) - (K_1K_2)^2$ denotes the determinant of the covariance matrix, S. Then, let us consider two random column vectors extracted from a 3-dimensional MHD, $\vec{x} = (x_1, x_2, x_3)^T$ and $\vec{y} = (y_1, y_2, y_3)^T$ such that $n = \sum_{i=1}^{3} x_i = \sum_{i=1}^{3} y_i$ where n is the number of sampled papers. Substituting Eq. (C.4) in Eq. (C.1), we write the square of the $d_{\mathcal{M}}$ between $\vec{x}$ and $\vec{y}$ as

$$\begin{aligned} d_{\mathcal{M}}(\vec{x}, \vec{y})^2 =& \frac{1}{\gamma\, det(\mathbf{S})} \begin{pmatrix} x_1 - y_1 & x_2 - y_2 \end{pmatrix} \begin{pmatrix} K_2(N - K_2) & +K_1K_2 \\ +K_1K_2 & K_1(N - K_1) \end{pmatrix} \begin{pmatrix} x_1 - y_1 \\ x_2 - y_2 \end{pmatrix} \\ =& \frac{1}{\gamma\, det(\mathbf{S})} \left\{ (x_1 - y_1)^2 K_2(N - K_2) + (x_2 - y_2)^2 K_1(N - K_1) \right. \\ & \left. + \, 2(x_1 - y_1)(x_2 - y_2)(K_1K_2) \right\} \\ =& \frac{1}{\gamma\, det(\mathbf{S})} \left\{ (x_1 - y_1)^2 K_2(K_1 + K_3) + (x_2 - y_2)^2 K_1(K_2 + K_3) \right. \\ & \left. + \, 2(x_1 - y_1)(x_2 - y_2)(K_1K_2) \right\} \\ =& \frac{1}{\gamma\, det(\mathbf{S})} \left\{ (x_1 - y_1)^2 K_2K_3 + (x_2 - y_2)^2 K_1K_3 \right. \\ & \left. + \, [(x_1 - y_1) + (x_2 - y_2)]^2 K_1K_2 \right\} \end{aligned}$$

where we have used $N = \sum_{i=1}^{3} K_i$. Recalling that $n = \sum_{i=1}^{3} x_i = \sum_{i=1}^{3} y_i$, we know that $(x_1 - y_1) + (x_2 - y_2) = (x_3 - y_3)$, so we write

$$d_{\mathcal{M}}(\vec{x}, \vec{y})^2 = \frac{1}{\gamma\, det(\mathbf{S})} \left\{ (x_1 - y_1)^2 K_2 K_3 + (x_2 - y_2)^2 K_1 K_3 + (x_3 - y_3)^2 K_1 K_2 \right\} \tag{C.5}$$

Then, by using the relation $det(\mathbf{S}) = N \prod_{i=1}^{3} K_i$, we have:

$$d_{\mathcal{M}}(\vec{x}, \vec{y})^2 = \frac{1}{\gamma} \sum_{i=1}^{3} \frac{(x_i - y_i)^2}{N K_i}; \tag{C.6}$$

noticing from Eq. (C.2) that $\gamma = S_{i,i}/(K_i(N - K_i))$, we obtain

$$d_{\mathcal{M}}(\vec{x}, \vec{y})^2 = \sum_{i=1}^{3} \frac{(x_i - y_i)^2}{S_{ii}} \frac{K_i(N - K_i)}{N K_i} = \sum_{i=1}^{3} \frac{(x_i - y_i)^2}{S_{ii}} \left(1 - \frac{K_i}{N}\right) \tag{C.7}$$

Finally, if we choose one of the two vectors to contain the expected values, μ_i, we re-obtain Eq. (3.5) since $(x_i - \mu_i)^2/S_{ii} = z_i^2$. To be precise, the covariance matrix is not defined for $i = 3$, however the relation $\gamma = \sigma_3^2/(K_3(N - K_3))$ holds and therefore also the final result.

Using Mathematica or similar softwares, it is easy to prove analytically that Eq. (3.5) holds for small dimensions. We have verified it until 6 dimensions. Moreover, we have numerically tested this formula by calculating the $d_{\mathcal{M}}$'s between the ranking vectors of the indicators and the vector of expected values, $\vec{\mu}$, with two different alternative methods: (1) by using Eq. (C.1), i.e., by inverting the covariance matrix, and (2) by using the eigenvalue decomposition of the covariance matrix[1]. The results of the three methods were all compatible with each other up to 10 decimal digits. The advantage of using Eq. (3.5) is that we can calculate the $d_{\mathcal{M}}$ between two arbitrary vectors without dealing

[1]The matrix $\mathbf{S}$ is symmetric and it has maximal rank because it is the covariance matrix of a multivariate distribution. Therefore, we can diagonalize it, $\mathbf{S} = \mathbf{B}^{-1}\mathbf{D}\mathbf{B}$ where the columns of $\mathbf{B}$ form an orthonormal basis; we can also write $\mathbf{S}^{-1} = \mathbf{B}^{-1}\mathbf{D}^{-1}\mathbf{B}$. With this, we have $d^2_{\mathcal{M}}(\vec{x}, \vec{y}) = \sum_i^{F-1} c_i/\lambda_i$, where $\{\lambda_i\}$ are the eigenvalues of $\mathbf{S}$ and $\{c_i\}$ are the coordinates of $\vec{x} - \vec{y}$ in the basis which diagonalizes $\mathbf{S}$, i.e. $c_i = \sum_k^{F-1}(x_k - y_k)B^{-1}_{ki} = \sum_k^{F-1}(x_k - y_k)B_{ik}$ where the last equality comes from the orthonormality of $\mathbf{B}$ which implies $\mathbf{B}^{-1} = \mathbf{B}^T$.

with any (computationally slow) matrix inversion or diagonalization, and the number of needed calculations scales linearly with the number of dimensions. Importantly, Eq. (3.5) allows also to assess the individual contribution of each dimension (i.e. of each category) to the $d^2_{\mathcal{M}}$. To our best knowledge, we are the first ones to have derived such explicit formula for $d_{\mathcal{M}}$ when the covariance matrix and the random vectors come from a MHD.

Appendix D

List of journals and universities

Top-100 journals according to Google metrics

In table D.1 we report the top-100 journals according to Google metrics. The only six journal that we were not able to automatically match between the journal names reported in `MAG17` and this list are: "the lancet oncology", "nber working papers", "ieee conference on computer vision and pattern recognition, cvpr", "british medical journal", "the lancet neurology", "journal of materials chemistry a".

Journal name	Journal name
nature	nature reviews genetics
the new england journal of medicine	immunity
science	chemical communications
the lancet	nature neuroscience
chemical society reviews	european heart journal
cell	monthly notices of the royal astronomical soc.
journal of the american chemical society	advanced functional materials
advanced materials	scientific reports
pnas	cancer research
chemical reviews	nature reviews immunology
nature communications	arxiv quantum physics
jama	nature reviews molecular cell biology
physical review letters	science translational medicine
angewandte chemie international edition	neuroimage
nano letters	annals of internal medicine
journal of clinical oncology	journal of materials chemistry
nucleic acids research	nature physics
energy & environmental science	nature reviews cancer
acs nano	cell stem cell
nature genetics	diabetes care
arxiv high energy physics - experiment	cancer cell
journal of the american college of cardiology	nanoscale
nature materials	physical review b
arxiv mesoscale and nanoscale physics	clinical cancer research
plos one	hepatology
the lancet oncology	cell metabolism
arxiv cosmology and ...	molecular cell
circulation	ieee transactions on power electronics
arxiv high energy physics - phenomenology	journal of financial economics
nature medicine	nature climate change
nber working papers	biomaterials
journal of high energy physics	arxiv high energy physics - theory
the astrophysical journal	obstetrical & gynecological survey
arxiv materials science	acs applied materials & interfaces
ieee conference on cvpr	physics letters b
blood	clinical infectious diseases
nature biotechnology	the lancet neurology
nature nanotechnology	gut
the cochrane database of systematic reviews	nature immunology
accounts of chemical research	environmental science & technology
nature photonics	journal of hepatology
nature methods	journal of materials chemistry. a
british medical journal	annals of oncology
neuron	european urology
physical review d	arxiv information theory
gastroenterology	the journal of physical chemistry c
renewable and sustainable energy reviews	nature reviews neuroscience
the american economic review	ieee transactions on industrial electronics
the journal of clinical investigation	ieee transactions on pattern analysis and ...
arxiv computer vision and ...	nature chemistry

Table D.1: The top-100 journals according to Google metrics (retrieved on the 01/06/2018).

Top-100 universities in Computer Science

In table D.2, we report the top-100 Univ. analyzed by [38]. Out of this 100 Univ., we matched 81 names from `MAG16`. To match the strings we have performed only simple preprocessing, i.e. transformed the affiliation name in lower case and removed punctuation. For example, the strings "Univ. of California, Berkeley" was transformed in "Univ. of california berkeley". We have also manually matched the "Swiss Federal Institute of Technology Zurich" with "ETH Zurich".

Univ. name	Univ. name
Australian National Univ.	RWTH Aachen Univ.
Boston Univ.	*Rutgers, the State Univ. of New Jersey*
Brown Univ.	Saint Petersburg State Univ.
California Institute of Technology	Seoul National Univ.
Carnegie Mellon Univ.	Stanford Univ.
Columbia Univ.	*Technical Univ. of Munich*
Cornell Univ.	*Texas A&M Univ.*
Delft Univ. of Technology	*The Univ. of Queensland*
Duke Univ.	Tsinghua Univ.
Durham Univ.	Univ. College London
ETH Zurich	Univ. of Amsterdam
École Normale Supérieure	Univ. of Bristol
École Polytechnique	Univ. of British Columbia
École Polytechnique Fédérale de Lausanne	Univ. of California, Berkeley
Free Univ. of Berlin	Univ. of California, Davis
Georgia Institute of Technology	Univ. of California, Los Angeles
Harvard Univ.	Univ. of California, San Diego
Heidelberg Univ.	Univ. of California, San Francisco
Hong Kong Univ. of Science and Technology	Univ. of California, Santa Barbara
Humboldt Univ. of Berlin	Univ. of Cambridge
Imperial College London	Univ. of Chicago
Johns Hopkins Univ.	Univ. of Copenhagen
KU Leuven	Univ. of Edinburgh
Karolinska Institute	Univ. of Helsinki
King's College London	Univ. of Hong Kong
Kyoto Univ.	*Univ. of Illinois at Urbana-Champaign*
LMU Munich	Univ. of Manchester
Leiden Univ.	Univ. of Maryland, College Park
Lomonosov Moscow State Univ.	*Univ. of Massachusetts*
London Business School	Univ. of Melbourne
London School of Economics & Political Sc.	Univ. of Michigan
Massachusetts Institute of Technology	Univ. of Minnesota
Mayo Medical School	Univ. of North Carolina at Chapel Hill
McGill Univ.	Univ. of Oxford
Michigan State Univ.	Univ. of Pennsylvania
Monash Univ.	Univ. of Pittsburgh
Nanyang Technological Univ.	Univ. of Southern California
National Autonomous Univ. of Mexico	Univ. of Sydney
National Taiwan Univ.	*Univ. of São Paulo*
National Univ. of Singapore	Univ. of Texas at Austin
New York Univ.	Univ. of Tokyo
Northwestern Univ.	Univ. of Toronto
Ohio State Univ.	Univ. of Warwick
Panthéon-Sorbonne Univ. – Paris 1	Univ. of Washington
Paris-Sorbonne Univ. – Paris 4	*Univ. of Wisconsin-Madison*
Pasteur Institute	Uppsala Univ.
Peking Univ.	Utrecht Univ.
Pennsylvania State Univ.	*Wageningen Univ. and Research Center*
Princeton Univ.	Washington Univ. in St Louis
Purdue Univ.	Yale Univ.

Table D.2: The top-100 Univ. reported by [38]. We list them in alphabetical order and use an *italic* font for those universities that we have not matched.

Appendix E

Stability of scientists' productivity

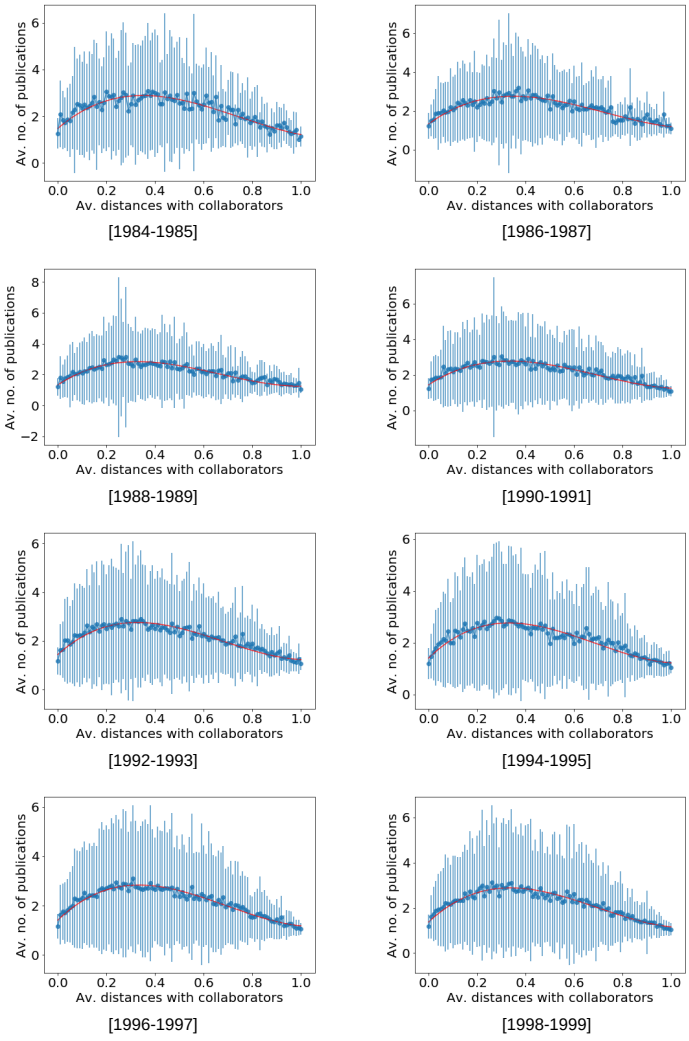


Figure E.1: Scientists productivity as defined in Chap. 6 between 1984 to 2000 using time windows of length 2. The red line is the fitted productivity using Eq. (7.1) and the `curve_fit` function of the `scipy` package.

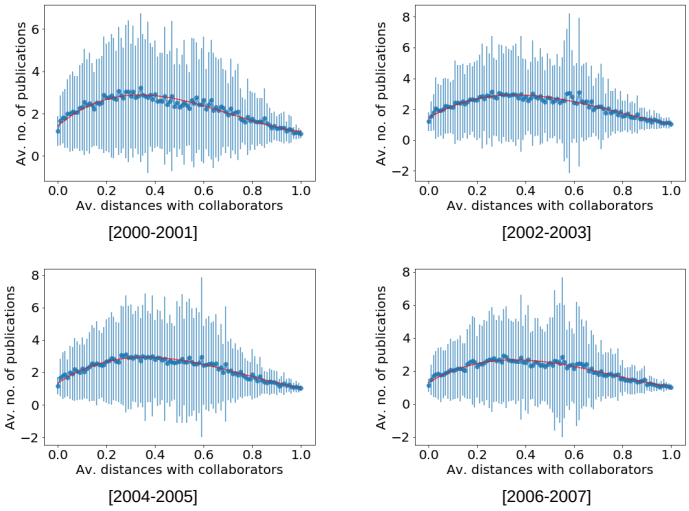


Figure E.2: Scientists productivity as defined in Chap. 6 between 2000 to 2006 using time windows of length 2. The red line is the fitted productivity using Eq. (7.1) and the `curve_fit` function of the `scipy` package.

List of Figures

List of Tables

Bibliography

[1] (2013). San Francisco Declaration on Research Assessment, `https://sfdora.org/`.

[2] Adams, J.; Gurney, K.; Jackson, L. (2008). Calibrating the zoom—A test of Zitt's hypothesis. *Scientometrics* **75(1)**, 81–95.

[3] Agrawal, A.; Kapur, D.; McHale, J.; Oettl, A. (2011). Brain drain or brain bank? The impact of skilled emigration on poor-country innovation. *Journal of Urban Economics* **69(1)**, 43–55.

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

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

[6] Albarrán, P.; Crespo, J. A.; Ortuño, I.; Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. *Scientometrics* **88(2)**, 385–397.

[7] Antonelli, C. (1996). Localized knowledge percolation processes and information networks. *Journal of Evolutionary Economics* **6(3)**, 281–295.

[8] Appelt, S.; van Beuzekom, B.; Galindo-Rueda, F.; de Pinho, R. (2015). Chapter 7 - Which Factors Influence the International Mobility of Research Scientists? In: A. Geuna (ed.), *Global Mobility of Research Scientists*, San Diego: Academic Press. pp. 177–213.

[9] Arenas, A.; Duch, J.; Fernández, A.; Gómez, S. (2007). Size reduction of complex networks preserving modularity. *New Journal of Physics* **9(6)**, 176.

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

[11] Bahar, D.; Hausmann, R.; Hidalgo, C. A. (2014). Neighbors and the evolution of the comparative advantage of nations: Evidence of international knowledge diffusion? *Journal of International Economics* **92(1)**, 111 – 123.

[12] Bahrami, B.; Olsen, K.; Latham, P. E.; Roepstorff, A.; Rees, G.; Frith, C. D. (2010). Optimally Interacting Minds. *Science* **329(5995)**, 1081–1085.

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

[14] Baringhaus, L.; Franz, C. (2004). On a new multivariate two-sample test. *Journal of multivariate analysis* **88(1)**, 190–206.

[15] Barrat, A.; Barthelemy, M.; Pastor-Satorras, R.; Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences* **101(11)**, 3747–3752.

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

[17] Baum, J. A.; Calabrese, T.; Silverman, B. S. (2000). Don't go it alone: Alliance network composition and startups' performance in Canadian biotechnology. *Strategic management journal* **21(3)**, 267–294.

[18] Becker, G. (1964). *Human capital: a theoretical and empirical analysis, with special reference to education*. National bureau of economic research publications: General series, Columbia University Press.

[19] Beechler, S.; Woodward, I. C. (2009). The global "war for talent". *Journal of International Management* **15(3)**, 273–285.

[20] Beine, M.; Docquier, F.; Rapoport, H. (2001). Brain Drain and Economic Growth: Theory and Evidence. *Journal of Development Economics* **64(1)**, 275–289.

[21] Bénassy, J. P.; Brezis, E. S. (2013). Brain drain and development traps. *Journal of Development Economics* **102**, 15–22.

[22] Bergstra, J.; Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research* **13(Feb)**, 281–305.

[23] Bergstrom, C. T.; West, J. D.; Wiseman, M. A. (2008). The Eigenfactor$^{\text{TM}}$ metrics. *Journal of Neuroscience* **28(45)**, 11433–11434.

[24] Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* **2008(10)**, P10008.

[25] Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology* **92(5)**, 1170–1182.

[26] Börner, K.; Maru, J. T.; Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences* **101(suppl 1)**, 5266–5273.

[27] Bornmann, L.; Daniel, H.-D. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation* **64(1)**, 45–80.

[28] Bornmann, L.; Daniel, H.-D. (2009). Universality of citation distributions–A validation of Radicchi et al.'s relative indicator cf= c/c0 at the micro level using data from chemistry. *Journal of the American Society for Information Science and Technology* **60(8)**, 1664–1670.

[29] Boucher, A.; Cerna, L. (2014). Current Policy Trends in Skilled Immigration Policy. *International Migration* **52(3)**, 21–25.

[30] Boyack, K. W.; Klavans, R.; Börner, K. (2005). Mapping the backbone of science. *Scientometrics* **64(3)**, 351–374.

[31] Brin, S.; Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* **30(1)**, 107–117.

[32] Burt, R. (1992). *Structural Holes: The Social Structure of Competition.* Harvard University Press.

[33] Butts, C. T. (2008). 4. A Relational Event Framework for Social Action. *Sociological Methodology* **38(1)**, 155–200.

[34] Butts, C. T. (2009). Revisiting the foundations of network analysis. *science* **325(5939)**, 414–416.

[35] Cañibano, C.; Otamendi, F. J.; Solís, F. (2011). International temporary mobility of researchers: a cross-discipline study. *Scientometrics* **89(2)**, 653–675.

[36] Chen, P.; Xie, H.; Maslov, S.; Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. *Journal of Informetrics* **1(1)**, 8–15.

[37] Chesbrough, H. W. (2006). The era of open innovation. *Managing innovation and change* **127(3)**, 34–41.

[38] Clauset, A.; Arbesman, S.; Larremore, D. B. (2015). Systematic inequality and hierarchy in faculty hiring networks. *Science advances* **1(1)**, e1400005.

[39] Coleman, J. S. (1988). Social Capital in the Creation of Human Capital. *American Journal of Sociology* **94**, S95–S120.

[40] Colliander, C.; Ahlgren, P. (2011). The effects and their stability of field normalization baseline on relative performance with respect to citation impact: A case study of 20 natural science departments. *Journal of Informetrics* **5(1)**, 101–113.

[41] Coscia, M.; Giannotti, F.; Pedreschi, D. (2011). A Classification for Community Discovery Methods in Complex Networks. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **4(5)**, 512–546.

[42] Cowan, R.; Jonard, N.; Ozman, M. (2004). Knowledge dynamics in a network industry. *Technological Forecasting and Social Change* **71(5)**, 469–484.

[43] Cowan, R.; Jonard, N.; Zimmermann, J. (2007). Bilateral collaboration and the emergence of innovation networks. *Management Science* **53(7)**, 1051–1067.

[44] Czaika, M.; Parsons, C. R. (2017). The gravity of high-skilled migration policies. *Demography* **54(2)**, 603–630.

[45] Danon, L.; Diaz-Guilera, A.; Duch, J.; Arenas, A. (2005). Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment* **2005(09)**, P09008.

[46] Das, T.; Teng, B. (2000). A resource-based theory of strategic alliances. *Journal of management* **26(1)**, 31–61.

[47] Davis, B. (1990). Reinforced random walk. *Probability Theory and Related Fields* **84(2)**, 203–229.

[48] De Domenico, M.; Omodei, E.; Arenas, A. (2016). Quantifying the diaspora of knowledge in the last century. *Applied Network Science* **1(1)**, 15.

[49] Deffuant, G.; Neau, D.; Amblard, F.; Weisbuch, G. (2000). Mixing beliefs among interacting agents. *Advances in Complex Systems* **3(4)**, 87–98.

[50] DeGroot, M. H. (1974). Reaching a consensus. *Journal of the American Statistical Association* **69(345)**, 118–121.

[51] Dolfsma, W. (2008). *Knowledge Economies: Organization, Location and Innovation*. Routledge Studies in Global Competition, Taylor & Francis.

[52] Eck, N. J. v.; Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of the American society for information science and technology* **60(8)**, 1635–1651.

[53] Ermann, L.; Frahm, K. M.; Shepelyansky, D. L. (2015). Google matrix analysis of directed networks. *Reviews of Modern Physics* **87(4)**, 1261–1310.

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

[55] Fernandez-Zubieta, A.; Geuna, A.; Lawson, C. (2015). Chapter 1 - What Do We Know of the Mobility of Research Scientists and of its Impact on Scientific Production. In: A. Geuna (ed.), *Global Mobility of Research Scientists*, San Diego: Academic Press. pp. 1–33.

[56] Fischer, M. M.; Fröhlich, J. (2001). *Knowledge, complexity and innovation systems*. Springer Science & Business Media.

[57] Fortunato, S. (2010). Community detection in graphs. *Physics reports* **486(3-5)**, 75–174.

[58] Fortunato, S.; Barthelemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences* **104(1)**, 36–41.

[59] Fortunato, S.; Bergstrom, C. T.; Börner, K.; Evans, J. A.; Helbing, D.; Milojević, S.; Petersen, A. M.; Radicchi, F.; Sinatra, R.; Uzzi, B.; *et al.* (2018). Science of science. *Science* **359(6379)**, eaao0185.

[60] Fortunato, S.; Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports* **659**, 1–44.

[61] Franceschet, M. (2011). PageRank: Standing on the shoulders of giants. *Communications of the ACM* **54(6)**, 92–101.

[62] Franzoni, C.; Scellato, G.; Stephan, P. (2014). The mover's advantage: The superior performance of migrant scientists. *Economics Letters* **122(1)**, 89–93.

[63] Franzoni, C.; Scellato, G.; Stephan, P. (2015). Chapter 2 - International Mobility of Research Scientists: Lessons from GlobSci. In: A. Geuna

(ed.), *Global Mobility of Research Scientists*, San Diego: Academic Press. pp. 35–65.

[64] Fruchterman, T.; Reingold, E. (1991). Graph Drawing by Force-directed Placement. *Software- Practice and Experience* **21(11)**, 1129–1164.

[65] Garas, A.; Tomasello, M. V.; Schweitzer, F. (2017). Newcomers vs. incumbents: How firms select their partners for R&D collaborations. *arXiv:1403.3298*.

[66] Garfield, E. (1964). "Science Citation Index"-A New Dimension in Indexing. *Science* **144(3619)**, 649–654.

[67] Garfield, E. (2006). Citation indexes for science. A new dimension in documentation through association of ideas. *International journal of epidemiology* **35(5)**, 1123–1127.

[68] Garfield, E. (2006). The history and meaning of the journal impact factor. *Journal of American Medical Association* **295(1)**, 90–93.

[69] Gargiulo, F.; Caen, A.; Lambiotte, R.; Carletti, T. (2016). The classical origin of modern mathematics. *EPJ Data Science* **5(26)**, 1–15.

[70] Georghiou, L. (1998). Global cooperation in research. *Research Policy* **27(6)**, 611 – 626.

[71] Geuna, A. (2015). *Global mobility of research scientists: The economics of who goes where and why.* Academic Press.

[72] Gibson, J.; McKenzie, D. (2014). Scientific mobility and knowledge networks in high emigration countries: Evidence from the Pacific. *Research Policy* **43(9)**, 1486 – 1495.

[73] Gilbert, N. (1997). A simulation of the structure of academic science. *Sociological Research Online* **2(2)**, 1–15.

[74] Gilbert, N. (2004). *Agent-based social simulation: dealing with complexity. Tech. rep.*, Center for Research on Social Simulation, University of Surrey, Guildford, UK.

[75] Gilbert, N.; Ahrweiler, P.; Pyka, A.; *et al.* (2014). *Simulating knowledge dynamics in innovation networks*. Springer.

[76] Gleich, D. F. (2015). PageRank beyond the Web. *SIAM Review* **57(3)**, 321–363.

[77] Gomes-Casseres, B.; Hagedoorn, J.; Jaffe, A. (2006). Do alliances promote knowledge flows? *Journal of Financial Economics* **80(1)**, 5–33.

[78] Gote, C.; Scholtes, I.; Schweitzer, F. (2019). git2net - Mining time-stamped co-editing networks from large git repositories. In: *Proceedings of the 16th International Conference on Mining Software Repositories*. MSR.

[79] Grant, R.; Baden-Fuller, C. (2004). A knowledge accessing theory of strategic alliances. *Journal of Management Studies* **41(1)**, 61–84.

[80] Groeber, P.; Schweitzer, F.; Press, K. (2009). How groups can foster consensus: The case of local cultures. *Journal of Artificial Societies and Social Simulation* **12(2)**, 1–12.

[81] Guimera, R.; Uzzi, B.; Spiro, J.; Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science* **308(5722)**, 697–702.

[82] Gulati, R. (1995). Social structure and alliance formation patterns: A longitudinal analysis. *Administrative science quarterly* **40(4)**, 619–652.

[83] Gulati, R.; Gargiulo, M. (1999). Where do interorganizational networks come from? *The American Journal of Sociology* **104(5)**, 1398–1438.

[84] Gulati, R.; Sytch, M.; Tatarynowicz, A. (2012). The rise and fall of small worlds: Exploring the dynamics of social structure. *Organization Science* **23(2)**, 449–471.

[85] Hagedoorn, J. (2002). Inter-firm R&D partnerships: an overview of major trends and patterns since 1960. *Research policy* **31(4)**, 477–492.

[86] Hagedoorn, J. (2003). Sharing intellectual property rights—an exploratory study of joint patenting amongst companies. *Industrial and Corporate Change* **12(5)**, 1035–1050.

[87] Hagedoorn, J.; Link, A. N.; Vonortas, N. S. (2000). Research partnerships. *Research Policy* **29(4-5)**, 567–586.

[88] Hanaki, N.; Nakajima, R.; Ogura, Y. (2010). The dynamics of R&D network in the IT industry. *Research policy* **39(3)**, 386–399.

[89] Hanneke, S.; Fu, W.; Xing, E. P. (2010). Discrete temporal models of social networks. *Electronic Journal of Statistics* **4**, 585–605.

[90] Harzing, A.-W.; Alakangas, S. (2017). Microsoft Academic: is the phoenix getting wings? *Scientometrics* **110(1)**, 371–383.

[91] Hayek, F. A. (1945). The use of knowledge in society. *The American economic review* **35(4)**, 519–530.

[92] Hegselmann, R.; Krause, U. (2002). Opinion dynamics and bounded confidence: models, analysis and simulation. *Journal of Artificial Societies and Social Simulation* **5(3)**, 1–33.

[93] Helbing, D.; Schweitzer, F.; Keltsch, J.; Molnár, P. (1997). Active walker model for the formation of human and animal trail systems. *Physical Review E* **56**, 2527–2539.

[94] Hicks, D.; Wouters, P.; Waltman, L.; De Rijcke, S.; Rafols, I. (2015). Bibliometrics: the Leiden Manifesto for research metrics. *Nature* **520**, 429–431.

[95] Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences* **102(46)**, 16569–16572.

[96] Howells, J. R. L. (2002). Tacit Knowledge, innovation and economic geography. *Urban Studies* **39(5-6)**, 871–884.

[97] Hug, S.; Ochsner, M.; Brandle, M. P. (2017). Citation analysis with microsoft academic. *Scientometrics* **111(1)**, 371–378.

[98] Hug, S. E.; Brändle, M. P. (2017). The coverage of Microsoft Academic: analyzing the publication output of a university. *Scientometrics* **113(3)**, 1551–1571.

[99] Inkpen, A. C.; Ross, J. (2001). Why do some strategic alliances persist beyond their useful life? *California Management Review* **44(1)**, 132–148.

[100] Jaccard, P. (1912). The distribution of the flora in the alpine zone. *New phytologist* **11(2)**, 37–50.

[101] Jaffe, A. B. (1986). Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits, and Market Value. *The American Economic Review* **76(5)**, 984–1001.

[102] Jaffe, A. B.; Trajtenberg, M. (2002). *Patents, citations, and innovations: A window on the knowledge economy.* MIT press.

[103] Jia, T.; Wang, D.; Szymanski, B. K. (2017). Quantifying patterns of research-interest evolution. *Nature Human Behaviour* **1(4)**, 0078.

[104] Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika* **18(1)**, 39–43.

[105] Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American documentation* **14(1)**, 10–25.

[106] Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* **46(5)**, 604–632.

[107] Kogut, B.; Zander, U. (1992). Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization science* **3(3)**, 383–397.

[108] Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari* **4**, 83–91.

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

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

[111] Lambiotte, R.; Rosvall, M.; Scholtes, I. (2019). From networks to optimal higher-order models of complex systems. *Nature Physics* **15(4)**, 313–320.

[112] Lancichinetti, A.; Fortunato, S. (2009). Community detection algorithms: A comparative analysis. *Physical Review E* **80**, 056117.

[113] Larivière, V.; Gingras, Y.; Sugimoto, C. R.; Tsou, A. (2015). Team size matters: Collaboration and scientific impact since 1900. *Journal of the Association for Information Science and Technology* **66(7)**, 1323–1332.

[114] Law, A. M.; Kelton, W. D. (1991). *Simulation modeling and analysis.* McGraw-Hill Inc.

[115] Le Pair, C. (1980). Switching between academic disciplines in universities in the Netherlands. *Scientometrics* **2(3)**, 177–191.

[116] Lee, B. K.; Lessler, J.; Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine* **29(3)**, 337–346.

[117] Leifeld, P.; Cranmer, S. J. (2019). A theoretical and empirical comparison of the temporal exponential random graph model and the stochastic actor-oriented model. *Network Science* **7(1)**, 20–51.

[118] Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly of applied mathematics* **2(2)**, 164–168.

[119] Leydesdorff, L. (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology* **58(9)**, 1303–1319.

[120] Leydesdorff, L.; Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the Association for Information Science and Technology* **60(2)**, 348–362.

[121] Liebeskind, J. P. (1996). Knowledge, Strategy, and the Theory of the Firm. *Strategic Managment Journal* **17**, 93–109.

[122] Lissoni, F.; Mairesse, J.; Montobbio, F.; Pezzoni, M. (2011). Scientific productivity and academic promotion: a study on French and Italian physicists. *Industrial and Corporate Change* **20(1)**, 253–294.

[123] Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences* **16(12)**, 317–323.

[124] Lundberg, J. (2007). Lifting the crown—citation z-score. *Journal of Informetrics* **1(2)**, 145–154.

[125] Luukkonen, T.; Persson, O.; Sivertsen, G. (1992). Understanding patterns of international scientific collaboration. *Science, Technology & Human Values* **17(1)**, 101–126.

[126] Ma, A.; Mondragón, R. J.; Latora, V. (2015). Anatomy of funded research in science. *Proceedings of the National Academy of Sciences* **112(48)**, 14760–14765.

[127] Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Science of India* , 49–55.

[128] Mahroum, S. (2000). Scientific Mobility. *Science Communication* **21(4)**, 367–378.

[129] Mariani, M. S.; Medo, M.; Zhang, Y.-C. (2015). Ranking nodes in growing networks: When PageRank fails. *Scientific Reports* **5**.

[130] Mariani, M. S.; Medo, M.; Zhang, Y.-C. (2016). Identification of milestone papers through time-balanced network centrality. *Journal of Informetrics* **10(4)**, 1207–1223.

[131] Marquardt, D. W. (1963). An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics* **11(2)**, 431–441.

[132] Marshakova, I. V. (1973). System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2 – Informatsionnye Protsessy I Sistemy.*

[133] Maslov, S.; Redner, S. (2008). Promise and pitfalls of extending Google's PageRank algorithm to citation networks. *The Journal of Neuroscience* **28(44)**, 11103–11105.

[134] McAllister, P. R.; Narin, F.; Corrigan, J. G. (1983). Programmatic evaluation and comparison based on standardized citation scores. *IEEE Transactions on Engineering Management* **(4)**, 205–211.

[135] McCain, K. W. (1991). Mapping economics through the journal literature: An experiment in journal cocitation analysis. *Journal of the American Society for Information Science* **42(4)**, 290–296.

[136] Medo, M.; Cimini, G.; Gualdi, S. (2011). Temporal effects in the growth of networks. *Physical Review Letters* **107**, 238701.

[137] Mincer, J. (1958). Investment in Human Capital and Personal Income Distribution. *Journal of Political Economy* **66(4)**, 281–302.

[138] Morgenstern, D. (2001). Proof of a conjecture by Walter Deuber concerning the distances between points of two types in Rd. *Discrete Mathematics* **226(1)**, 347–349.

[139] Mowery, D.; Oxley, J.; Silverman, B. (1998). Technological overlap and interfirm cooperation: implications for the resource-based view of the firm. *Research Policy* **27(5)**, 507–523.

[140] Nanumyan, V. (2014). Master Equation for Hetereogenouse Evolution of Interdependent Networks.

[141] Nanumyan, V. (2018). Structure and Dynamics of Collaborative Knowledge Networks.

[142] Narin, F. (1991). Globalization of research, scholarly information, and patents–ten year trends. *The Serials Librarian* **21(2-3)**, 33–44.

[143] Newman, M. (2003). The Structure and Function of Complex Networks. *SIAM review* **45(2)**, 167–256.

[144] Newman, M. (2010). *Networks: an introduction.* Oxford University Press.

[145] Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* **98(2)**, 404–409.

[146] Newman, M. E. (2009). The first-mover advantage in scientific publication. *Europhysics Letters* **86(6)**, 68001.

[147] Newman, M. E. J. (2001). Who is the best connected scientist? A study of scientific coauthorship networks. *Physical Review E* **64**, 016132.

[148] Newman, M. E. J. (2002). Assortative Mixing in Networks. *Physical Review Letters* **89(20)**, 208701.

[149] Newman, M. E. J. (2004). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences* **101(suppl 1)**, 5200–5205.

[150] Newman, M. E. J.; Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E* **69**, 026113.

[151] Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization science* **5(1)**, 14–37.

[152] Nonaka, I.; Takeuchi, H. (1995). *The Knowledge-creating Company: How Japanese Companies Create the Dynamics of Innovation.* Oxford University Press.

[153] Nooteboom, B. (1999). *Inter-firm Alliances: Analysis and Design.* Routledge.

[154] Nooteboom, B.; Van Haverbeke, W.; Duysters, G.; Gilsing, V.; Van den Oord, A. (2007). Optimal cognitive distance and absorptive capacity. *Research policy* **36(7)**, 1016–1034.

[155] Noyons, E. C.; van Raan, A. F. (1998). Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research. *Journal of the American society for information science* **49(1)**, 68–81.

[156] Olesen, J. M.; Bascompte, J.; Dupont, Y. L.; Jordano, P. (2007). The modularity of pollination networks. *Proceedings of the National Academy of Sciences* **104(50)**, 19891–19896.

[157] Owen-Smith, J.; Powell, W. W. (2004). Knowledge networks as channels and conduits: The effects of spillovers in the Boston biotechnology community. *Organization science* **15(1)**, 5–21.

[158] Pan, R. K.; Sinha, S.; Kaski, K.; Saramäki, J. (2012). The evolution of interdisciplinarity in physics research. *Scientific reports* **2**, 551.

[159] Parolo, P. D. B.; Pan, R. K.; Ghosh, R.; Huberman, B. A.; Kaski, K.; Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics* **9(4)**, 734–745.

[160] Perra, N.; Goncalves, B.; Pastor-Satorras, R.; Vespignani, A. (2012). Activity driven modeling of time varying networks. *Scientific Reports* **2**, 469.

[161] Petersen, A. M. (2015). Quantifying the impact of weak, strong, and super ties in scientific careers. *Proceedings of the National Academy of Sciences* **112(34)**, E4671–E4680.

[162] Petersen, A. M.; Riccaboni, M.; Stanley, H. E.; Pammolli, F. (2012). Persistence and uncertainty in the academic career. *Proceedings of the National Academy of Sciences* **109(14)**, 5213–5218.

[163] Phelps, C. (2003). *Technological exploration: A longitudinal study of the role of recombinatory search and social capital in alliance networks.* Ph.D. thesis, New York University, Graduate School of Business Administration.

[164] Pinski, G.; Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management* **12(5)**, 297–312.

[165] Podolny, J. M. (1993). A status-based model of market competition. *American journal of sociology* **98(4)**, 829–872.

[166] Polanyi, M. (1966). *The tacit dimension*. University of Chicago Press.

[167] Porter, M. A.; Onnela, J.-P.; Mucha, P. J. (2009). Communities in networks. *Notices of the American Mathematical Society* **56(9)**, 1082–1097.

[168] Powell, W.; Koput, K.; Smith-Doerr, L. (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative science quarterly* **41(1)**, 116–145.

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

[170] Price, D. d. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* **27(5)**, 292–306.

[171] Price, D. J. d. S. (1951). Quantitative measures of the development of science. *Archive Internationale d'Histoire de Sciences* **4(14)**, 86–93.

[172] Pyka, A.; Fagiolo, G. (2007). *Agent-based modelling: a methodology for neo-Schumpeterian economics*, Elgar companion to neo-schumpeterian economics.

[173] Pyka, A.; Gilbert, N.; Ahrweiler, P. (2006). Simulating Knowledge-Generation and -Distribution Processes in Innovation Collaborations and Networks. Discussion Paper Series.

[174] Pyka, A.; Gilbert, N.; Ahrweiler, P. (2009). Agent-based modelling of innovation networks–the fairytale of spillover. In: *Innovation networks*, Springer. pp. 101–126.

[175] Pyka, A.; Windrum, P. (2000). *The Self-Organisation of Innovation Networks. Research Memorandum 020*, Maastricht University, Maastricht Economic Research Institute on Innovation and Technology (MERIT).

[176] Radicchi, F.; Castellano, C. (2011). Rescaling citations of publications in physics. *Physical Review E* **83(4)**, 046116.

[177] Radicchi, F.; Castellano, C. (2012). Testing the fairness of citation indicators for comparison across scientific domains: The case of fractional citation counts. *Journal of Informetrics* **6(1)**, 121–130.

[178] Radicchi, F.; Castellano, C. (2012). Why Sirtes's claims (Sirtes,2012) do not square with reality. *Journal of Informetrics* **6(4)**, 615–618.

[179] Radicchi, F.; Fortunato, S.; Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences* **105(45)**, 17268–17272.

[180] Ramasco, J. J.; Dorogovtsev, S. N.; Pastor-Satorras, R. (2004). Self-organization of collaboration networks. *Physical Review E* **70(3)**, 036106.

[181] Raub, W.; Weesie, J. (1990). Reputation and efficiency in social interactions: An example of network effects. *American Journal of Sociology* **96(3)**, 626–654.

[182] Rokach, L.; Maimon, O. (2005). Top-down induction of decision trees classifiers-a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* **35(4)**, 476–487.

[183] Romer, P. M. (1994). The Origins of Endogenous Growth. *The Journal of Economic Perspectives* **8(1)**, 3–22.

[184] Rosenkopf, L.; Almeida, P. (2003). Overcoming local search through alliances and mobility. *Management science* **49(6)**, 751–766.

[185] Rosenkopf, L.; Nerkar, A. (2001). Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry. *Strategic Management Journal* **22(4)**, 287–306.

[186] Rosenkopf, L.; Padula, G. (2008). Investigating the microstructure of network evolution: Alliance formation in the mobile communications industry. *Organization Science* **19(5)**, 669.

[187] Rosenkopf, L.; Schilling, M. (2007). Comparing alliance network structure across industries: observations and explanations. *Strategic Entrepreneurship Journal* **1(3-4)**, 191–209.

[188] Rosvall, M.; Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences* **105(4)**, 1118–1123.

[189] Rosvall, M.; Bergstrom, C. T. (2010). Mapping change in large networks. *PloS one* **5(1)**, e8694.

[190] Rosvall, M.; Esquivel, A. V.; Lancichinetti, A.; West, J. D.; Lambiotte, R. (2014). Memory in network flows and its effects on spreading dynamics and community detection. *Nature communications* **5**, 4630.

[191] Ružička, M. (1958). Anwendung mathematisch-statistischer Methoden in der Geobotanik (Synthetische Bearbeitung von Aufnahmen). *Biológia* **13**, 647–661.

[192] Sampson, R. C. (2007). R&D alliances and firm performance: the impact of technological diversity and alliance organization on innovation. *Academy of Management Journal* **50(2)**, 364–386.

[193] Sarigöl, E.; Garcia, D.; Scholtes, I.; Schweitzer, F. (2017). Quantifying the effect of editor–author relations on manuscript handling times. *Scientometrics* **113(1)**, 609–631.

[194] Sarigöl, E.; Pfitzner, R.; Scholtes, I.; Garas, A.; Schweitzer, F. (2014). Predicting scientific success based on coauthorship networks. *EPJ Data Science* **3(1)**, 1.

[195] Saxenian, A. (2005). From Brain Drain to Brain Circulation: Transnational Communities and Regional Upgrading in India and China. *Studies in Comparative International Development* **40(2)**, 35–61.

[196] Scellato, G.; Franzoni, C.; Stephan, P. (2017). A mobility boost for research. *Science* **356(6339)**, 694–697.

[197] Scharnhorst, A.; Börner, K.; van den Besselaar, P. (2012). *Models of Science Dynamics: Encounters Between Complexity Theory and Information Sciences*. Understanding Complex Systems, Springer Berlin Heidelberg.

[198] Schilling, M. A. (2009). Understanding the alliance data. *Strategic Management Journal* **30(3)**, 233–260.

[199] Schilling, M. A.; Steensma, H. K. (2001). The use of modular organizational forms: An industry-level analysis. *Academy of management journal* **44(6)**, 1149–1168.

[200] Schmoch, U. (2008). Concept of a technology classification for country comparisons. *Final report to the world intellectual property organisation, WIPO*.

[201] Scholtes, I. (2017). When is a network a network? Multi-order graphical model selection in pathways and temporal networks. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 1037–1046.

[202] Scholtes, I.; Pfitzner, R.; Schweitzer, F. (2014). The Social Dimension of Information Ranking: A Discussion of Research Challenges and Approaches. In: *Socioinformatics – The Social Impact of Interactions between Humans and IT*. pp. 45–61.

[203] Scholtes, I.; Wider, N.; Garas, A. (2016). Higher-order aggregate networks in the analysis of temporal networks: Path structures and centralities. *European Physical Journal B* **89(3)**, 1–15.

[204] Scholtes, I.; Wider, N.; Pfitzner, R.; Garas, A.; Tessone, C. J.; Schweitzer, F. (2014). Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks. *Nature communications* **5**, 5024.

[205] Schubert, A.; Braun, T. (1986). Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics* **9(5-6)**, 281–291.

[206] Schubert, A.; Telcs, A. (2014). A note on the Jaccardized Czekanowski similarity index. *Scientometrics* **98(2)**, 1397–1399.

[207] Schultz, T. W. (1961). Investment in human capital. *The American Economic Review* **51(1)**, 1–17.

[208] Schweitzer, F.; Behera, L. (2009). Nonlinear voter models: the transition from invasion to coexistence. *The European Physical Journal B-Condensed Matter and Complex Systems* **67(3)**, 301–318.

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

[210] Shiffrin, R. M.; Börner, K. (2004). Mapping knowledge domains.

[211] Simon, H. (1957). *Models of Man.* John Wiley.

[212] Sinatra, R.; Wang, D.; Deville, P.; Song, C.; Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science* **354(6312)**, aaf5239.

[213] Sinha, A.; Shen, Z.; Song, Y.; Ma, H.; Eide, D.; Hsu, B.-j. P.; Wang, K. (2015). An overview of microsoft academic service (mas) and applications. In: *Proceedings of the 24th International Conference on World Wide Web*. pp. 243–246.

[214] Sirtes, D. (2012). Finding the Easter eggs hidden by oneself: Why Radicchi and Castellano's (2012) fairness test for citation indicators is not fair. *Journal of Informetrics* **6(3)**, 448–450.

[215] Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science* **24(4)**, 265–269.

[216] Small, H. G.; Koenig, M. E. (1977). Journal clustering using a bibliographic coupling method. *Information Processing & Management* **13(5)**, 277–288.

[217] Snijders, T. A.; Van de Bunt, G. G.; Steglich, C. E. (2010). Introduction to stochastic actor-based models for network dynamics. *Social networks* **32(1)**, 44–60.

[218] Stanley, H. E. (1968). Dependence of critical properties on dimensionality of spins. *Physical Review Letters* **20(12)**, 589.

[219] Sun, X.; Kaur, J.; Milojević, S.; Flammini, A.; Menczer, F. (2013). Social dynamics of science. *Scientific reports* **3**, 1069.

[220] Tang, J.; Zhang, J.; Yao, L.; Li, J.; Zhang, L.; Su, Z. (2008). Arnetminer: extraction and mining of academic social networks. In: *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. pp. 990–998.

[221] Thomson-Reuters (2013). SDC Platinum dataset, `http://thomsonreuters.com/sdc-platinum/`.

[222] Tomasello, M. V.; Burkholz, R.; Schweitzer, F. (2017). Modeling the formation of R&D alliances: an agent-based model with empirical validation. *Economics-ejournal* **2017(107)**, 1–18.

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

[224] Tomasello, M. V.; Perra, N.; Tessone, C. J.; Karsai, M.; Schweitzer, F. (2014). The Role of Endogenous and Exogenous Mechanisms in the Formation of R&D Networks. *Scientific Reports* **4**, 5679.

[225] Tomasello, M. V.; Tessone, C. J.; Schweitzer, F. (2015). The effect of R&D collaborations on firms' technological positions. In: *Proceedings of the 10th International Forum IFKAD 2015*.

[226] Tomasello, M. V.; Tessone, C. J.; Schweitzer, F. (2016). A model of dynamic rewiring and knowledge exchange in R&D networks. *Advances in Complex Systems* **19(01n02)**, 1650004.

[227] Tomasello, M. V.; Vaccario, G.; Schweitzer, F. (2017). Data-driven modeling of collaboration networks: a cross-domain analysis. *EPJ Data Science* **6(1)**, 22.

[228] Torvik, V. I. (2015). MapAffil: A Bibliographic Tool for Mapping Author Affiliation Strings to Cities and Their Geocodes Worldwide. *D-Lib Magazine* **21(11/12)**.

[229] Torvik, V. I.; Smalheiser, N. R. (2009). Author name disambiguation in MEDLINE. *ACM Transactions on Knowledge Discovery from Data* **3(3)**, 11.

[230] Uzzi, B.; Mukherjee, S.; Stringer, M.; Jones, B. (2013). Atypical combinations and scientific impact. *Science* **342(6157)**, 468–472.

[231] Vaccario, G.; Luca, V.; Schweitzer, F. (2018). Reproducing scientists' mobility: A data-driven model. *arXiv:1811.07229*.

[232] Vaccario, G.; Medo, M.; Wider, N.; Mariani, M. S. (2017). Quantifying and suppressing ranking bias in a large citation network. *Journal of informetrics* **11(3)**, 766–782.

[233] Vaccario, G.; Tomasello, M. V.; Tessone, C. J.; Schweitzer, F. (2018). Quantifying knowledge exchange in R&D networks: A data-driven model. *Journal of Evolutionary Economics* **28(3)**, 461–493.

[234] Vaccario, G.; Verginer, L.; Schotles, I. (2019). The empirical flow of knowledge. *(Working paper)*.

[235] Van Eck, N. J.; Waltman, L. (2008). Appropriate similarity measures for author co-citation analysis. *Journal of the American Society for Information Science and Technology* **59(10)**, 1653–1661.

[236] Van Leeuwen, T. N.; Medina, C. C. (2012). Redefining the field of economics: Improving field normalization for the application of bibliometric techniques in the field of economics. *Research Evaluation* **21(1)**, 61–70.

[237] Verginer, L.; Riccaboni, M. (2018). Brain-circulation network: The global mobility of the life scientists. *(Working Paper)*.

[238] Verginer, L.; Riccaboni, M. (2018). *The mobility of Life Scientists – Patterns and Determinants*. Ph.D. thesis, IMT School for Advanced Studies Lucca.

[239] Veugelers, R.; Bouwel, L. V. (2015). Chapter 8 - Destinations of Mobile European Researchers: Europe versus the United States. In: A. Geuna (ed.), *Global Mobility of Research Scientists*, San Diego: Academic Press. pp. 215–237.

[240] Vinkler, P. (1986). Evaluation of some methods for the relative assessment of scientific publications. *Scientometrics* **10(3-4)**, 157–177.

[241] Vlachý, J. (1981). Mobility in physics. *Czechoslovak Journal of Physics B* **31(6)**, 669–674.

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

[243] Wagner, C. S.; Whetsell, T. A.; Mukherjee, S. (2019). International research collaboration: Novelty, conventionality, and atypicality in knowledge recombination. *Research Policy* **48(5)**, 1260–1270.

[244] Walker, D.; Xie, H.; Yan, K.-K.; Maslov, S. (2007). Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment* **2007(06)**, P06010.

[245] Walker, G.; Kogut, B.; Shan, W. (1997). Social Capital, Structural Holes and the Formation of an Industry Network. *Organization Science* **8(2)**, 109–125.

[246] Walsh, J. R. (1935). Capital concept applied to man. *The Quarterly Journal of Economics* **49(2)**, 255–285.

[247] Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics* **10(2)**, 365–391.

[248] Waltman, L.; van Eck, N. J. (2010). The relation between eigenfactor, audience factor, and influence weight. *Journal of the American Society for Information Science and Technology* **61(7)**, 1476–1486.

[249] Waltman, L.; van Eck, N. J.; van Raan, A. F. (2012). Universality of citation distributions revisited. *Journal of the American Society for Information Science and Technology* **63(1)**, 72–77.

[250] Waltman, L.; Yan, E.; van Eck, N. J. (2011). A recursive field-normalized bibliometric performance indicator: An application to the field of library and information science. *Scientometrics* **89(1)**, 301–314.

[251] Wang, D.; Song, C.; Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science* **342(6154)**, 127–132.

[252] Wasserman, S.; Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*. *Psychometrika* **61(3)**, 401–425.

[253] Wernerfelt, B. (1984). A resource-based view of the firm. *Strategic Management Journal* **5(2)**, 171–180.

[254] White, H. D.; Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the American Society for information Science* **32(3)**, 163–171.

[255] Wu, L.; Wang, D.; Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature* **566(7744)**, 378–382.

[256] Wuchty, S.; Jones, B. F.; Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. *Science* **316(5827)**, 1036–1039.

[257] Xu, J.; Wickramarathne, T. L.; Chawla, N. V. (2016). Representing higher-order dependencies in networks. *Science advances* **2(5)**, e1600028.

[258] Yao, L.; Wei, T.; Zeng, A.; Fan, Y.; Di, Z. (2014). Ranking scientific publications: the effect of nonlinearity. *Scientific Reports* **4**, 6663.

[259] Zhang, Z.; Cheng, Y.; Liu, N. C. (2014). Comparison of the effect of mean-based method and z-score for field normalization of citations at the level of Web of Science subject categories. *Scientometrics* **101(3)**, 1679–1693.

[260] Zhou, J.; Zeng, A.; Fan, Y.; Di, Z. (2016). Ranking scientific publications with similarity-preferential mechanism. *Scientometrics* **106(2)**, 805–816.

[261] Zitt, M.; Ramanana-Rahary, S.; Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics* **63(2)**, 373–401.

[262] Zweig, K. A. (2011). Good versus optimal: Why network analytic methods need more systematic evaluation. *Central European Journal of Computer Science* **1(1)**, 137–153.