Shortest paths to success: Network indicators of performance in innovation ecosystems

Moreno Bonaventura

A thesis submitted to the University of London for the degree of
Doctor of Philosophy

School of Mathematical Sciences and School of Business and Management
Queen Mary, University of London
United Kingdom

September 2016

to those who believe in me and to those who tell me "*it's impossible*",
they are both equally valuable

Abstract

In this thesis I show how various theories and methodologies borrowed from complexity science, organisation science, and network science can be suitably integrated to provide a comprehensive and interdisciplinary approach to the study of innovation processes. I study the network foundations of success in innovation ecosystems and I conduct several empirical investigations to identify those network characteristics that are expected to correlate with positive outcomes and success. I assess the extent to which the diversity and the strength in the networks of relationships boost the performance and success of scientists and early-stage firms. To this end I analyse two large-scale data sets about scientific publishing and start-up firms by making use of already existing topological network measures and by proposing novel measures to characterise the degree of interdisciplinarity and access to diverse pools of knowledge in scientific collaborations. Results provide empirical support to the idea that collaboration sustains innovation and performance by facilitating knowledge diffusion, acquisition and creation. First, results indicate that the networks of interaction between start-ups have a strong impact on the firms' long-term success. Second I find that, while abandoning specialisation in favour of moderate degrees of interdisciplinarity deteriorates scientific performance, very interdisciplinary scientists tend to outperform specialised ones. Additionally, I address the computational challenges related to the size of the data sets used and their time-varying nature. In particular I focus on the scalability challenges of incremental graph algorithms. The thesis contributes in this direction by proposing new efficient algorithms and data struc-

tures to handle and to analyse large graphs whose nodes and edges change rapidly over time. These efforts have been collected and made available to the public in the form of a web platform (http://lab.startup-network.org/) and an open-source python package, NetworkL (https://networkl.github.io/).

Acknowledgments

This thesis would have not been possible without the direct and indirect contribution of many people. In the lines that follow I wish to express my gratitude for the support, the opportunities and the encouragement I have received. First of all, I shall thank my supervisors Vito and Pietro for having given me a life-changing opportunity. I am honoured for the guidance they have given me during the last four years. Under their advice I have strengthened myself both personally and professionally and I am grateful for the trust and independence they have given me in setting up my research agenda. I especially thank Vito for having embraced me in his group and for making me feel part of the family. Since the very beginning I have felt welcomed, included, and supported. I especially thank Pietro for his great dedication in having taken care of my professional development. I have appreciated his effort in overcoming the difference in our professional backgrounds and I thank him for having provided me with precious new knowledge. Equally, I wish to thank the person who I regard as a hero of science: Enzo. From explaining the R-J vectors on the board of the lab in Catania, to sketching formulas on the desks of the Queen Mary's common room, his deep knowledge, truthful manners, and openness have been always of great inspiration to me. Sometimes my research activity has gone beyond the campus of Queen Mary. I wish to thank all the people involved in the Startup-Network project. Mario and Luciano for having given me great trust in leading the most risky part of our project; Andrea, Marco, and Salvo for having pushed the project with enthusiasm and great energy. All the guys in Paradigma Innovation for

having supported our efforts not just with technology but also with friendship. I wish to thank all the great people of the BEST consortium for the time they have devoted to the project. It has been a great honour to collaborate with them. I wish to thank Jim Webber from Neo Tehcnology and James Cameron from Accel Partners for their time, support and advice. I want to thank all the people in the maths department at Queen Mary University of London for the warm welcome I have received since the very beginning of my PhD. I wish to thank all PhD colleagues with whom I have shared many years of work and in particular Andrea (the King of machines), Jacopo (the Boss) and Federico. A very big thank you to all the people that have made my life in the UK great and pleasant: Massimo and his kitchen phase transition, Andrea La Rosa and his multiple house moves, the house hunt of Peppe Marino, and the python coding, guitar playing and singing with Laura, Marci and Cami. Many thanks go to Alex, not just for having read this thesis with patience, but also for stepping into my life so graciously. I have infinite gratitude for Valerio and Nils with whom I have built and shared a very special home and unique experience. Equally, I thank all the friends and special people of my life who are now spread across the world: Chiara, Alice, Tiziana, Mariachiara, Giorgia, Claudia, Vincenzo, Roberto, Nicola, Gregorio, Michele and many others. Last but not least, with great emotion, I wish to thank my family; Tella, Marcello and Sofia. No matter where I am in the world I know there exists only one shortest path to success, and that is the path towards home.

Table of Contents

List of Abbreviations

WWS network	WorldWide Startup network
APS	American Physical Society
WOS	Web of Science
PACS	The Physics and Astronomy Classification Scheme
SGM	Sparse Geodesic Matrix
BSGM	Biconnected Sparse Geodesic Matrix
EU	European Commission
EVCE	European Private Equity & Venture Capital Association

Chapter 1

Introduction

What characterises extremely successful innovative companies? What makes researchers able to produce scientific works with exceptional impact? The recently emerged fields of computational social science and science of success offer new theoretical and methodological perspectives to tackle these questions. The abundance of digital data allows to track the dynamics and performance of human-made systems at unprecedented level of details and offers today the unique opportunity to investigate empirically the complex dynamics responsible for the creation of innovation and new knowledge. In this thesis I show that large-scale digital data and a social-network prospective make it exceptionally easy to identify the early signals of success for innovative business and scientific activities. The results of my empirical endeavour cast a new light on the current understanding of innovation ecosystems and have the potential to transform the way institutions and corporations monitor, control, and optimise innovation processes.

This work lies at the intersection between several disciplines: computer science, math-

ematics, sociology, and creativity studies. On the one hand the thesis includes the first empirical investigation of the largest available data sets about innovation and start-up companies. On the other hand the methodology and analysis draw extensively on hypotheses and theories from the social sciences. In recent years, encouraged by the increasing availability of digital data, social sciences and computer science have converged in the so-called *computational social science*: a discipline which attempts to model and understand human behaviours and outcomes of socio-economic systems. Studies in the context of computational social science have proliferated in recent years and have covered a variety of domains. Among them, particular interest has been raised by the so called *science of success* whose goal is the identification of patterns and regularities which characterise and anticipate various forms of success and positive outcomes. Here, I focus the attention on the success of *innovation ecosystems* and their actors: start-up founders, investors, and scientists. To this end I conduct an empirical study of the dynamics of innovation ecosystems (e.g., social interactions, collaboration, exchange of ideas, creation of start-ups, production of research articles) which are responsible for the generation of innovation, knowledge and societal advances. Institutions and large organisations are currently devoting substantial attention to innovation. Governments are keen to promote science, innovation and entrepreneurship because they are the drivers for technological advances, job creation, economic growth, and societal transformations. A recent review by the MIT Sloan School of Management has highlighted how established organisations view innovation as critical to corporate success [1]. The review reports the words by William Ford Jr., CEO of Ford Motor Co., announcing that *"innovation will be the compass by which the company sets its direction"*, Jeffrey Immelt, CEO of General

Electric Co., talking about the *"Innovation Imperative, a belief that innovation is central to the success of a company"*, and Steve Ballmer, CEO of Microsoft Corp., stating that *"Innovation is the only way that Microsoft can keep customers happy and competitors at bay"*.

However, the push towards innovation does not come without a cost. The exploration of unanticipated solutions entail great uncertainty, repeated failures, and sunk costs. After all, the fields of science and business are ripe with stories of failures which have led to accidental breakthroughs: penicillin, X-rays, microwave, graphene, to name only a few. History suggests that great discoveries are inherently intertwined with chance and that the success of ambitious researches or radically new business models is hardly predictable. Mistakes in anticipating successful innovations abound in history. A remarkable example, concerning the first attempts to develop jet engines, is reported in the Jane Jacobs' book *The Economy of Cities*. The author writes: "...in 1937, when the jet airplane engine had already been developed in Britain, a committee of distinguished aeronautical experts in the United States, to whom this event was not yet known, having studied the possibilities of jet propulsion, came to the conclusion that it was not practicable. It was their recommendation that attempts to develop jet propulsion were dropped". Similarly, in the mid-1980s, McKinsey & Co. estimated that the overall worldwide potential for the cellular-telephone market was 900,000 units, missing by several order of magnitudes the current number of mobile-phone daily subscribers [2]. Given the important role that innovation plays in our society and the hidden costs due to its inherent unpredictability, a question then arise naturally: how can one mitigate failure and risk in innovation processes? Scholars have investigated the process of innovation and knowledge creation

for decades. Social scientists, urbanists, and economists have explored the multiple facets of innovation and its determinants. A common theme, agreed by many, is that innovation and new knowledge come from the exchange and recombination of already existing knowledge and ideas [3–5]. Even though it is largely recognised that knowledge exchanges primarily occur through social interactions, the lack of data about various forms of human interaction and collaboration has given researchers little opportunity to empirically investigate the social dimension of innovation. My overarching research hypothesis is that the network of collaboration among scientists on the one hand, and professional interactions among start-up companies on the other, are the conduits facilitating the knowledge and information transfer within innovation ecosystems. Access to relevant knowledge and information is crucial for the achievement of research objectives or business goals, and the access to resources and opportunities strongly depends on the actor's position in the network and on its immediate contacts [6, 7]. Therefore, the research idea developed in this thesis is that, by directly analysing data on microscopic social interactions, I will be able to unveil early signals of long-term performance of both scientists and startups. In doing so I aim to define a practical tool to improve the efficiency of innovation ecosystems, mitigate risk, and sustain the efforts of innovators around the world. I will move towards this direction equipped with a combination of approaches borrowed from network science, complexity theory, computer science, graph theory, algorithms and social science. The computational challenges in analysing large-scale and time-varying graphs also offer the opportunity to revise current approaches to the manipulation and analysis of network data sets. I contribute in this direction by proposing advanced computational strategies suitable for time-resolved graphs and by

introducing the python package *NetworkL* which I have developed and made available to the research community.

1.1 Outline of the thesis

Chapter 2 illustrates the general context in which the empirical analysis presented in the thesis unfold. In the first section I review the literature on the *science of success*, a recently emerged field of research aimed at exploring patterns that govern the path to success. The second section is devoted to the description of different innovation processes and practices (e.g., scientific, technological, product, and business innovations) and the common mechanisms underlying the creation of innovation in the areas of science and business. I will highlight the crucial role that knowledge flows, diversity, social dynamics, and social networks have in fostering innovation in both areas. The anatomy of innovation ecosystems and the role of their different actors (e.g., universities, start-ups, venture fund, governments) are discussed. I will also highlight the recent shift from closed innovation to open innovation and how big corporations are changing the way to renovate their business models and maintain competitive advantage. The last section illustrates the benefit of a network approach in mapping and describing the landscape of innovation ecosystems, and finally in identifying the determinants of their success.

Chapter 3 highlights the role that science and start-up firms have in innovation processes and sheds light on the intimate relationship between research activities and the life in a technological start-up. I will present two case studies which will lead us to describe in details the elementary processes occurring in innovation ecosystems (e.g., exchange

and recombination of ideas, knowledge transfer, collaborations patterns) and to better define the hypothesis at the root of the empirical analysis presented in Chapters 4 and 5.

Chapter 4 presents the results of various empirical investigations conducted on two data sets about start-up companies. In the first two sections I provide a detailed description of the data sets and the methodology used to construct the World Wide Start-up (WWS) network. The third section illustrate a methodology to predict the success of start-ups companies from the network of their interactions (the WWS network). The methodology, the performances of the method, and robustness checks are discussed in this section while additional analysis of confounding factors are presented in Appendix B. In Appendix B.2 I propose a methodology to characterise various start-up ecosystems at the level of cities, and to outline their differences and similarities in a quantitative way. In particular the analysis is focused on the subgraph of interactions between companies and people within the same city. From each subgraph I construct the *fingerprint* of the local innovation ecosystem and I then compare the fingerprints of various cities. Results show a clear distinction in the patterns of interaction of European and US innovation communities.

Chapter 5 presents the results of various empirical investigations conducted on two data sets about scientific publishing. In particular I propose a method to characterise the degree of interdisciplinarity of a scientist (personal interdisciplinarity), as well as the scientists' exposure to variegate knowledge through his/her collaborators (social interdisciplinarity). I then study the impact of this two types of interdisciplinarity on scientific success. To measure success I use a sophisticated citation-based measures which account for variations in: (i) patterns and volume of citations across sub-fields and disciplines;

(ii) attractiveness of research topics over time; and (iii) the starting year and duration of authors' careers. Result show that both specialised and interdisciplinary scientists can be successful; yet extreme interdisciplinarity provides competitive advantage over extreme specialisation.

Chapter 6 describes NetworkL, a python package for the longitudinal analysis of time-varying complex networks which I have developed. In the first section I present a brief introduction to incremental graph problems, and a review of the relevant literature. The second section illustrates the main contribution of the library, i.e., the sparse biconnected geodesic matrix (SBGM). The third section describe the methods, functions, and variables implemented in the NetworkL package, and how to use them. The fourth section present some benchmarking tests and show the potential of NetworkL to save up 70% of memory on real-world network data sets. Last section is devoted to the future road map.

Chapter 7 is devoted to conclusions, discussion, and directions for future work.

Chapter 2

Success in innovation ecosystems

This chapter illustrates the socio-economic context and the research field in which the empirical analysis presented in the thesis unfolds. In the first section I shall present an introduction to the *science of success* and an overview on the most recent literature in this field. The second section is devoted to the description of different innovation processes and practices (e.g., scientific, technological, product, and business innovations) and the common mechanisms underlying the creation of innovation in the areas of science and business. I will highlight the crucial role that knowledge flows, diversity, social dynamics, and social networks have in fostering innovation in both areas. The anatomy of innovation ecosystems and the role of their different actors (e.g., universities, start-ups, venture fund, governments) are discussed. I will also highlight the recent shift from closed to open innovation and how big corporation are changing the way to renovate their business models and maintain competitive advantage. Section 2.3 discusses the factors that drive success and failures in innovation processes and illustrates the main

hypotheses tested later in the empirical analysis of Chapters 4 and 5. Lastly, Section 2.4 describes the benefit of a network approach in mapping and describing the landscape of innovation ecosystems, and identifying the determinants of their success.

2.1 What is science of success?

"science of success" is a term recently emerged to refer to a broad field of research at the intersection between social sciences, computer science, statistics, mathematics and engineering which mainly concerns with the identification of patterns and regularities in large-scale electronic data sets to characterise and anticipate various form of success and positive outcomes. The scope of applicability is enormous: over the last few years scholars is this field have covered a number of domains and have investigated the performances and success of athletes [8–14], researchers and scientific articles [15–20], universities [21], teams [22, 23], inventors and patents [24, 25], start-up companies [26], online petitioning [27], crowd-funding campaigns [28], cultural objects such as movies, books, and music [29], as well as the popularity of video on the Web, hash-tags on Twitter [30], the effectiveness of viral marketing campaign [31], and many others.

In the domain of sports, scholars have investigated historical data about baseball matches in the US [9], tested the effectiveness of training strategies in cycling by using data collected through smart-phone apps [10]. Other works have drawn upon the great level of details at which data on football matches is collected by new camera technologies [11]. By analysing the ball's trajectory and the players' movements across the pitch, authors have been able to characterise and distinguish quantitatively the various playing

styles and strategies imposed by the coaches [11]. Interestingly, the method proposed was able to distinguish the changes in a team's overall playing after the replacement of the coach across multiple football seasons and correlate the playing strategy with the team performance. Additionally, digitalised data on tennis matches has allowed scientists to device methodologies to anticipate the best tennis players one year in advance [32] or untangle professional performances from fame and popularity [33]. The study of the pattern of success in sport data is not only attractive for researchers and appealing for the public. It also provides significative competitive advantage to betting companies which, through the data, are able to optimise their odds. In general, virtually any business of the new millennium can improve its performances by means of data-driven strategies. Data is often regarded as the New Oil[1], an extremely valuable asset thanks to which companies can: improve efficiency, optimise supply chains, manage the availability of products on shelves, understand the habits of current or potential customers, leverage on social marketing, target advertising with extreme precisions, optimise transportation and deliveries, improve dynamic pricing, know, track and improve the performance of employees.

Marketing based on social influence and data-driven management of human resources are two fields which have recently found equally intense interest both from academia and industry. Aral has extensively investigated information diffusion on large-scale social networks and its impact on social contagion in viral marketing campaigns [34, 34–37]. Pentland and collaborators at MIT's Human Dynamics Laboratory have studied the performance of teams in a variety of industries, and identified those aspects which account

[1]For instance, Ann Winblad, investor and senior partner at Hummer-Winblad Venture, mentioned the sentence "Data is the new oil" during the CNBC TV-show denominated "The Pulse of Silicon Valley".

for the success and failure of team-based projects [22]. The researchers used wearable electronic sensors that collected data on the social behaviour of team members. Results showed that the most important predictor of a team's success was its communication pattern. It is worth noticing that both Aral's and Pentland's studies places a strong emphasis on the social dimension of performance and success.

Researchers across several disciplines have focused on the study of scientific performance by drawing on bibliometrics techniques and the large availability of data on scholarly publications, in the context of what is called the *"science of science"* [16–19]. Since citations are often used as criterion to award grants, or rank applicants competing for academic positions, virtually all the studies on academic performances are based on some citations-based measures. Indeed, the number of citations received is largely regarded as an indication of the relevance and quality of a paper as well as of its authors prestige and scientific success. In [19] a study has been conducted on the temporal profiles of the number of citations received by individual articles in different domains and unveiled an universal dynamics in the citations patterns which can be used to predict, at the early-stage, the ultimate impact of a given article. In [38] the role of geography and movements of academics across institutions, and their relations to career performances have been investigated. Uzzi et al. have highlighted the increasing dominance of multi-authored articles and studied the advantage that larger collaborations provide in terms of citations [39]. In [18] it has also been shown that scientific impact increases when atypical combination of knowledge is injected into conventional combinations of knowledge from prior works. Ma et. al. have used social network analysis techniques to investigate the anatomy of funded research [21]. In particular, it has been

highlighted the presence of cohesive and dense core in the network between universities collaborating on funded research projects. Belongings to the network core appeared to be crucial both for attracting additional financial resources and producing high quality research. Funding bodies themselves have drawn on the abundance of digital data to inform systematically their investment process. For instance, the EU commission has recently adopted the *Innovation Radar*, a semi-automatic data-driven tool to identify founded research project with high innovation potential. In [6] it has been shown how certain collaboration patterns and positions occupied within the co-authorship networks impact on scientists' performances. The authors have investigated the extent to which several centrality indices (degree, eigenvector, betweenness centrality [40], k-core [41–43]), measured on the network of collaboration between scientists, are able to predict whether an article will be highly cited after publication.

Advances in science rarely arise from individual and isolated contributions [39]. Even single-authored articles rely on the knowledge produced by others which can be gained by reading articles, attending conferences, and interact with other researchers. As a result, scientific activities, like other human activities, are not free from the influences of social aspects. Social dynamics, e.g, trust, reciprocity, power and authority, social cognition and social information filtering, shape the social dimension of the evolution of science [44–46]. Social barriers such as distrust, envy, fear, prevent the free flow of information, knowledge, discourage collaboration, consume individuals' energy, and ultimately reduce the opportunities for scientific advances and discoveries. By limiting the opportunity for recombination of ideas, social barriers are expected to reduce the chances of achieving not only scientific advances, but also, in general, all forms of innovation (technologies,

inventions, business models). Indeed, the human dimension and the social interactions have been placed at the core of Saxenian's comparative analysis between the Boston and the Silicon Valley areas [47], and are essential in Horowitt and Hwang's explanation of the great success achieved in Silicon Valley [48].

As data becomes ubiquitous, more accessible, and pervasive in our lives, we are now in a better position to understand and predict the outcomes of virtually any human activity (sports, science, collaboration, business). In this vein my work focuses on understanding the success of innovative activities by combining approaches from the social sciences [49] and network science [50] with large-scale longitudinal data sets on scientific production and start-up firms.

To summarise, science of success is a modern interdisciplinary area of investigation at the intersection between academia and industry. Nowadays, the abundance of digital data gives to decision makers within companies, institutions and funding bodies the opportunity to radically change the way decisions are made and resources invested. Insights from the data have the potential to help design better business strategies, promote positive behavioural and cultural changes and optimise investments. However, without a robust and general methodological framework to predict, from the data, future outcomes in a variety of different domains, the opportunity to accelerate societal advances and growth remains unexpressed.

2.2 Innovative activities: types, mechanisms and trajectories

Universities, research centres, R&D departments are traditionally regarded as the places where innovation and novel ideas are generated. Our society would not have been transformed by great technological inventions such as the transistor, the Internet, the Web, without the passionate joint effort of researchers and scholars from industry and academia. One of the greatest human achievements to date, the man landing on the Moon surface, was the result of an exceptional allocation of resources, assembly of talents, and advances in technologies. In 1969 the *"small step for [a] man, one giant leap for mankind"* has attracted the attention of about 600 million people on Earth and indirectly transformed their lives (e.g., through the positive impact of space exploration on satellite communication, navigation systems, whether forecast, and knowledge of the universe). In comparison, in 2015 the platform Facebook, created by one single college student with significantly fewer resources than was the case with NASA, has attracted the attention of nearly 1.59 billion active users. Assessed against the world's population Facebook has achieved a penetration in humans' lives 1.37 times greater than the one achieved during the live-cast of the Apollo11 mission. Yet, the level of ambition of the two enterprises remains disproportionate. These two opposite examples synthesise how the anatomy of innovation processes has radically changed over the last half century: from investments in long-term "Big Problems", such as landing a man on the moon, to short-term money-makers such as Facebook [51]. In what follows, I will illustrate this shift from a more traditional form of technological innovation, based on the development

of entirely new tools or materials, characterised by high complexity of knowledge and specialised expertise, to more recent business-driven innovations processes, which draw mainly on the recombination of existing technology, involve less complex and more accessible knowledge, occur faster, are more disruptive, and sometimes emerge from collective and distributed contributions.

The anatomy of modern innovation. The reason why innovation is commonly linked to universities, research centres and laboratories is that it is typically associated with scientific or technical discoveries which emerge from a delicate mix of highly complex knowledge, expensive materials or tools, and specialised competence. To master the intrinsic complexity of scientific or technical advances long-term training and substantial fundings are required. These constitute a strong entry barrier and limit the number of people that, at a given time, can contribute to a specific innovation process or initiative. Governmental policies or business visions guide significantly the trajectories of innovation process by setting goals, directions, and of course by allocating funding strategically. However, innovation processes in the technological domain have no predetermined duration and require continuos iteration over several steps (research, development, demonstration, production and deployment). As a result significant societal transformations usually happen after decades since the inception of a certain innovation strategy [2].

Despite technology pervades modern society, highly complex technical advances do not account for all forms of innovations. In fact, the most disruptive innovations of the last decades are those that simply recombine already existing technologies and less com-

[2] As an example think about the ongoing research on quantum computation, and the notable funding resources allocated by the European Commission.

plex knowledge, into new form of tools, products, services or business models [51]. There are two combined reasons for which innovation processes based only on the recombination of existing resources occur faster and are more disruptive than other more conventional types of innovation. First, the building blocks of technology are becoming more and more available to a wider audience. As the cost to access the building blocks of technology reduces, the number of people potentially available to contribute and collaborate in defining the next innovations increases [4]. As an example, the use of electronic boards and programming languages, once only in the hands of specialised engineers, are nowadays taught even at high-schools [3]. This two pieces of technology constitute, alone, the basis of announced forthcoming innovations, such as the Internet of Things (IOT). Even the current most visionary scientific endeavour, the one occurring in Geneva at the *Conseil Européen pour la Recherche Nucléaire* (CERN), can nowadays benefit from the wide contribution of the collectivity. The very same *Arduino* boards adopted as teaching tools in high-schools, are used at CERN to collect data from the Large Hadron Collider (LHC), making even school kids potentially able to contribute to the research. Moreover, machine-learning experts around the world have been given the opportunity to contribute to the Higgs Boson hunting by joining an online challenge hosted on the kaggle.com website. Secondly, lower cost and lower knowledge complexity means also more chances to iterate the experimentation, development, and diffusion cycles. It is well know that successful innovations are not the outcome of the very first attempt. Trial-and-error iterations underlie most innovations, and the lower are the costs and the complexity of each iteration, the more iterations can be performed with a given amount of resources and time. As a result, the larger the number of iterations, the higher the

[3] see for example the Arduino electronic board, `http://www.arduino.cc/` and the Scratch programming language developed at MIT `https://scratch.mit.edu/`

chances that a successful result will be achieved before the resources and the motivation of the innovators are used up.

The outputs of innovations that carry potential economic value become rapidly linked to business activities. Innovations find their way to market and impact on society often in conjunction with customers' needs and ambitious business visions. Innovations produced by the digital Era (e.g., Google) are the greatest example of recombination-based innovations that find rapidly a way to markets. In this categories we include smart-phone or Web-based Apps that try to transform and facilitate every aspect of life: access to information, communication, collaboration in working environments, entertainment, transportation, friendship. The global connectivity gives entrepreneurs from any part of the world the opportunity to reach rapidly a larger audience with new digital products and services. As an example, a modest investment of a few thousands euros has allowed the young Italian start-up *Ganiza* to reach and acquire more that 50,000 users for its *social planning* mobile application. The drop in the cost of customer acquisition through the Web has important consequences. Newly born companies which provides better products and services for the masses have more chances to overcome established competitors with solid customer base and rip away their clients. Massive and rapid shifts in the usage of online digital products are not unusual. The recent and sudden collapse of the Microsoft's MySpace platform in favour of other social network platforms shows that the competitive advantage of first movers in the digital domain is not as crucial as in other sectors. Given the low entry barriers, and the relatively simple knowledge required for the realisation of digital products, virtually any computer science student has attempted to emulate the Zuckerberg's success of Facebook.

However, not all disruptive innovations with a direct business opportunity have exclusively a digital form. One example is the recent explosion of interest in drones and quadcopters: small to medium-size robots with flying capabilities and equipped with sensors, cameras, or even weapons, accordingly to their size and engine power. The reason why their development and adoption have become so widespread only in the last few years is linked to a series of contingent factors, not all widely known. First, and more importantly, their autonomous flying capability relies heavily upon high-precision accelerometers: those small electronic chips embedded in smart-phones which enable the screen to rotate accordingly to the landscape/portrait orientation. Because of the economies of scale associated with the huge production of smart-phones, high-precision accelerometers have been made available on the market for ridiculous prices. Additionally, the previously mentioned programmable electronic boards allow to connect easily accelerometers to propellers of different sizes and powers, and provide the basic ground for increasingly richer flight logics and capabilities. The result of such a variegate combination of technologies is that almost anyone in the world is provided with a low-cost access to all the tools necessary to explore the opportunities hidden behind autonomous flying robots. The emergence of advanced flight capabilities has drawn deeply on the collective aspect of modern innovation. The variety of solutions available today in the drones market would not have been achievable if based on the effort of a few companies only. Instead, numbers of developers and hobbyists all over the world have created and shared a multitude of computer algorithms to track and follow a moving object in space, navigate GPS tracks, optimise stabilisation, fly in closed environment, return to charging bases and many others. Each feature can be linked to some specific applications: from

entertainment and fun to surveillance, from 3D mapping of the world surface to futuristic delivery services [4]. Some applications translate so directly into business opportunities that even venture funds dedicated exclusively to drones have emerged [5]. One of the sectors where quadcopters have had a breaking impact is the one of professional aerial filming. Equipped with 6 propellers or more to sustain high-resolution cameras, they make possible to realise professional aerial filming without the need of expensive helicopters and with the advantages of unique camera stabilisation and great control over the shots perspective. A home-made filming quadcopter, carried in an small backpack, serves faithfully independent film-makers even in the most remote and hostile shooting conditions, and open them the doors to compete against the larger budgets of established movie studios.

The above examples, from Moon landing to Facebook, research at CERN to school kids, and from movie studios to independent film makers, can offer a better sense of how innovation processes have undergone fundamental changes over the last century. To summarise, the recombination of technologies and ideas has tremendously accelerated over the last few decades thanks to a rapid drop in the cost of information exchange, access to knowledge, and mobility of talents. Innovation has become a more open phenomenon which involves the entire collectivity. The distinction between production and consumption of innovation has become less pronounced. Indeed by consuming innovations, individual can tap new opportunities to further innovate. Moreover, a number of distinct contributions of many people around the world can be recombined more easily into new products and services that can have rapid impact on the entire collectivity.

[4] Amazon is one of the companies which has announced experimental drones delivery `http://www.amazon.com/b?ie=UTF8&node=8037720011`

[5] See `https://angel.co/drone-vc`

The personal computers of a few decades ago and the more recent smart-phones, mobile Apps, 3D printers or programmable electronic boards make anyone capable of rethinking and reshaping our technological future. Global connectivity fosters collaboration and serendipity, and allows disruptive innovations to spring up everywhere in the world, spread rapidly, and create new business opportunities.

Innovation and established organisations. As the pace at which innovation is generated increases, new challenges for established institution (i.e. mature corporations, universities, research centres) arise. The multitude of people and small organisations betting on innovative activities represent a threat to incumbents. In fact there is no internal R&D spending that can compete with the global crowd of individuals and the speed at which they moves, innovate and produce new products that can disrupt established business. As a result, firms have started to, or have been forced to, open up their boundaries and draw opportunities for innovation from the collectivity. One very illustrative example is concerned with the telecommunication domain. We see every day an incremental and steady enhancement in mobile technologies. The speed of mobile internet connectivity has rapidly moved from the GPRS to the 4G speed which allows mobile users to get access to a great variety of contents. Yet, the investments in technology supported by telecommunication companies is destined to capitalise less and less. The more the network of communication opens up to the Internet the more telco companies lose their advantage point and their control on the revenue and the value generated on their infrastructures. It is well understood that the contents of the information generated and flowing on the tele-communication networks have more value than the value of communication services itself. In other words the value of transferring 1MB of data

is lower than the value of having access to, or controlling, the content of the data. The giant telco companies have their original revenue model, based only on providing the infrastructure on which communication occurs, at risk. Company such as WhatsApp or Viber have shown how rapidly changing technological regimes can destroy almost entirely large portions of revenue streams (i.e. those of short-message systems (SMS)). Even simple technologies can have enormous impact and companies which do not foreseen technological transformations and timely incorporate them in their business model faces substantial risks.

Large and established corporations have understood that, in order to remain competitive, must outsource part of their innovation processes to the collectivity. The process of harvesting ideas that lie beyond the formal boundaries of the company goes under the name of *open innovation* [1, 2]. If on the one hand the implementation of open innovation process brings considerable costs such us those arising from the resolution of intellectual property ownership issues or the lack of trust between the parties, on the other hand open innovation provides a company with access to a vast pool of ideas, much greater then the one available internally. Evidence of the interest for open innovation is the growing number of challenges, *call for ideas*, incubators and accelerators, promoted by big companies such as Airbus[6], Enel Energy[7], Microsoft[8], Samsung[9], Google[10], Telecom Italia[11], and many others[12]. Public competitions are precisely meant to explore the *collective mind* [52] provided by the community in search of new ideas, solutions

[6] https://www.airbus-fyi.com/challenges
[7] http://lab.enel.com/
[8] https://www.imaginecup.com
[9] https://developer.samsung.com/events/developer-challenge
[10] http://www.google.org/global-impact-awards/challenge/
[11] http://www.workingcapital.telecomitalia.it
[12] http://spacex.com

and products. Sometimes new opportunities come from individuals, more often they are enclosed in young companies such as start-ups. Large companies facilitate the birth and growth of fragile ideas also with incubators spaces: shared offices where start-ups can connect with mentors, advisors, customers, investors, but also other companies. As stated by Bill Ford, CEO of the homonymous automobile manufacturer, during the 2015 Web Summit in Dublin:

"This [the start-up incubators] is all new for us but we HAVE to do it. If i think the world we are about to enter into we are going to need partnership with today technology companies and with start-ups [...] I do think partnership is going to be important. If you think back to my great grandfather where the vision was to have a completely vertically integrated company where they made everything except the tires [...] that model would be very different in the future. It would be a series of partnership because one company can't and probably shouldn't know it all, or do it all."

Tons of start-ups, together with their patents and technologies, are acquired every day by large corporations. In this way the corporation secures new assets, reduces opportunities for competitors and lowers significantly the costs of R&D and internal innovation processes. The collectivity can outperform R&D departments not because R&D teams are less skilled than random individuals drawn from public competition, but simply because outside of the firms boundaries there is an enormous, and nowadays more easily accessible, pool of knowledge and solutions to draw from. Additionally, as stated earlier, innovation that involves less complex knowledge entails smaller costs and may initiate

faster transformation processes. Established corporations are therefore motivated to support those processes, practises, and initiatives that can stimulate the generation of start-ups.

2.3 What drives innovation, success and failure

The push towards innovation has intrigued scholars for years. Since the exploration of unknown and unanticipated solutions entails great uncertainty and sunk costs, much research has focused on the investigation of factors driving innovation processes, the reason why some attempts at innovating fail, and the way in which the intrinsic risks can be mitigated. Several traits of successful innovation mechanisms have been singled out and are presented in what follows.

Creativity and recombination of ideas play a crucial role in innovation process. While they are not sufficient, alone, to give birth to rapidly growing innovation clusters such as the Silicon Valley, or to bring scientific discovery to life, it is largely recognised that they constitute essential ingredients of innovation [3–5]. Therefore, scholars in the field of innovation have devoted substantial effort to understand which mechanisms favour or hinder creativity and successful recombination processes [7, 18, 25, 53–60]. Even though the perfect recipe to build flourishing innovation ecosystems does not exist, a number of studies have reached a fairly clear understanding of which aspects can stimulate the generation of innovation [4, 61–63]. There are two main *pillars* on which the argument unfolds:

- Recombination, creativity and the birth of novel ideas are processes which occur

within an individual's mind. However, the majority of the steps required in the attempt to generate novel and successful ideas involve an intense interaction with the outside world and other individuals. For this reason the locus of innovation has often been identified in the interaction between individuals rather than within the individual [64].

- The paths and the intermediate steps which lead to successful innovation are rarely foreseeable. Innovation processes are often associated with randomness, fortuity, coincidence, and serendipity [65]. Building innovation is a heuristic exploration of unknown and unanticipated solutions and resembles more a trial-and-error process than the execution of a predetermined plan [66–69].

Even though it is hard to engineer and directly control such processes, it is still possible to adopt certain interventions to create a fertile soil on which these processes are expected to occur at a high rate. In other words, where and when a novel and successful idea will be created remains largely unpredictable. Nevertheless, the probability that a certain innovation will emerge within a given time and space can be estimated and properly adjusted through suitable interventions and policies. To understand how innovation emerges I list below three main *principles* involved in the processes of recombination and creativity.

1) Understanding the context. Innovators need the ability to understand the context in which their work takes place. Analysing trends, expectations, customer needs, or current research issues and challenges, is crucial for entrepreneurs, scientists, and decision-makers in such a way that they can identify the right targets and concentrate

their efforts on the right paths. Moreover a proper understanding of the economy, society and science necessarily draws on the integration of multiple perspectives leading to collectively held interpretations and meanings. People that fail to account for others' perspectives are likely to be trapped in their biased personal picture of the world which reflects reality only partially. Innovation attempts starting from fallacious premises and biased assumptions are unlikely to be successful and to generate technological, scientific or societal transformations.

2) Failing fast. As the first attempts are very likely to be unsuccessful, a good practice is to invest resources on innovation strategies with a wider scope and with multiple exit options. This allows to reuse the outputs, results, and experience from previous efforts to sustain necessary deviations from unfruitful paths. A breadth-first exploration of the innovation space, in which resources are allocated to different but interrelated trials, is more sustainable that a deep-first exploration which allocates all resources to a unique, well-defined and rigid long-term plan. Breadth-first exploration allows to fail and disprove hypotheses at a rapid pace so as to leave resources available when a promising path is eventually found.

3) Drawing on variegate knowledge. The ability to draw on a great amount and variety of ideas, knowledge, and the ability to master it, substantially increase the opportunity to gather distinct pieces of information and pools of knowledge that can eventually recombine in successful ways. However, since no single individual can hold all the knowledge produced by the humanity, the process of accumulation of variegate knowledge must be conducted strategically and socially. While innovators can benefit from a global perspective over all knowledge domains, they should also have the ability to distinguish

combination of pieces of knowledge that stick well together from those that do not. In this discerning, collaborations with experts are crucial as they provide fast access to knowledge details without the associated costs. In some sense collaboration further reduces the cost of the breadth-first exploration of the innovation space as it provides a deeper view on certain paths of exploration and can help foreseen unfruitful ones.

The two *pillars* and the three *principles* mentioned above depict innovation as a socially-aided heuristic search process in which all goals are not the end of the search but are themselves hypotheses which need to be supported or disproved [69–71]. The knowledge gained at each intermediate goal is used to rule out unfruitful paths and to start again the search with different and more clear directions. An important aspect which emerges from my description is the social dimensions of innovation. In fact collaborators have a crucial role in reducing the efforts to search for, and access, relevant knowledge, and to build a reliable context perspective. Additionally, the various pieces of knowledge and materials put together in novel combinations by one single individual, often are derived from external experiences and from the interactions with others.

The *pillars* and *principles* I have condensed and described in the previous paragraphs offer ideas for research on innovation and can inspire a number of research hypotheses. **Steps 1) and 3) constitute the basis for the hypotheses tested in the empirical investigations presented in Chapter 4 and Chapter 5.** Here these hypotheses have been presented at a conceptual level while, within chapters 4 and 5, they will be discussed in more detail and turned into measurable quantities. In particular, in Chapter 4 I will test whether the centrality in the network of interactions between firms increases their chance of long term success. The network centrality can be indeed regarded as a proxy

for the potential access to knowledge and opportunities through immediate contacts. More central firms are expected to have greater and easier access to creative inputs and diverse pools of knowledge. In Chapter 5 I will focus more directly on knowledge categories and, to account for the principles 3), I will investigate whether or not pursuing interdisciplinary research benefits scientists careers.

The social dimension of innovation Collaboration, being a social construct can be affected by a number of social dynamics. The social network in which an individual is embedded significantly determines the amount and variety of perspectives, knowledge, and advice to sustain the trial-and-error process [54, 72]. Closer and denser social circles are expected to provide the same information, advice and knowledge on a certain problem or obstacle [62]. Resources from our closer peers are indeed likely to be redundant and, in order to find novelty, innovators have to search far from their neighbours. However, this process can be costly and unfruitful because, as we abandon our social circle, the level of trust and openness to collaboration decreases. For this reason, cultural transformations which promote collaboration among strangers for the sake of long-term mutual gain, against short term selfishness, have been regarded as the very fundamental ingredients of successful innovation ecosystems [48]. Two opposite aspects compete here. First, cultural phenomena typically diffuse in society through peer-to-peer interactions, the emergence of role models and practices, and the use of feedback mechanisms which discourage free-riding and penalise bad behaviour [48, 73, 74]. The more people interact and are densely connected, the easier is the diffusion of culture and the establishment of social norms which foster long-term collaborations. However, as the intensity of the interaction and collaboration increases the opportunity to access novelty decreases. Consequently, the

probability that different and distant perspectives and ideas will meet and recombine by chance reduces.

The trade-off between dense and sparse social interactions and their relative advantage have intrigued scholars for decades [72, 74–80]. The natural way to investigate quantitatively these aspects is to adopt a structural perspective and formalise interactions as a graph (or network). Social network scientists have proposed several measures (effective size, simmelian brokerage [73], efficiency [80], clustering [43]) to characterise the structural position of an individual in the network and distinguish between brokerage position (an actor interacting with mutually disconnected contacts) and cohesive structures (an actor interacting with mutually interconnected contacts). Various studies have shown that brokerage position provide competitive advantages in terms of salaries for managers [81], profit margin for companies [82], access to better job opportunities [83] However, mainly because of the lack of data, two aspects have been overlooked in the current literature: (i) the extent to which non-interacting contacts actually provide non-overlapping information and (ii) the role of global connectivity. Firstly, even though the absence of structural redundancy has been widely used as a proxy for non-redundancy of information, there is not guarantee that non interacting contacts actually provide diverse knowledge. In particular, since the creation of social connections is often driven by homophily, it is likely that the contacts of an actor are indeed similar, draw on the same sources of knowledge, or have similar backgrounds, even if they are not directly linked in the network. It is therefore important to assess directly the similarity and dissimilarity of actors in the network and the actual opportunities to access disparate knowledge by looking directly at properties and characteristics of the actors. Second,

looking only at the *local* connectivity, important information about the opportunity to access information may be missed. Two individuals, identical with respect to local connectivity patterns (e.g., same number of contacts and clustering coefficients) but located in central and peripheral regions of the global network may have radically different access to information. Overcoming this limitations is at the basis of the empirical analysis of Chapters 4 and 5 where the data sets considered include several nodes' properties and cover the entire systems under study, allowing a reconstruction of the global network.

2.4 Innovation, complexity, and the network approach

In the previous section I have shown that the social dimension of innovation emerges naturally from the dynamics that sustain knowledge recombination and creativity (sharing of ideas and resources, collaboration, trust, structural position in the graph of social relationships). In my view innovation can be regarded as a complex collective phenomenon, in which a multitude of agents jointly act to achieve goals and objectives (e.g., access, recombine and create knowledge), but whose global dynamics and outcomes (i.e., when and where a successful innovation will be generated) can be hardly explained and predicted simply in terms of individual agents' decisions and actions. The study of such kind of systems is precisely the scope of complexity science and network science, and in this section I will illustrate how the approaches borrowed from these theories are particularly suitable for the empirical investigations presented in Chapters 4 and 5. Additionally, I will highlight the intimate relation between complexity and network science and the reason at the basis of the growing popularity of graphs and networks for the description

of socio-economic systems. This will lead me to regard the network approach as an essential tool for the empirical investigation of the determinant of success in innovation ecosystems.

While reductionist approaches restrict the focus on the characteristics and properties of the individual components of a system, complexity theory aims at understanding a systems by investigating the relationships between its components [84]. Complexity theory has proven to be extremely successful in describing how rich behaviours of many biological, technological, and social systems, emerge from trivial dynamics of interrelated individual elements. Phenomena such as the functioning of the human brain, collective behaviour of flocks, congestions in transportation and communication systems, spread of disease information and gossip have been described and modelled by coupling relatively simple dynamics of elements with complex topologies of interactions. A extensive description of the wide applicability of complexity theory would be out of the scope of this thesis; comprehensive details can be found in the following books, reviews, and articles [50, 85–88].

I begin my discussion of complex networks by using as an example one of the most archetypal and intriguing complex systems studied: the human brain. The brain is made of several billion of neurones and of intricate electrochemical inter-connections that produce the most spectacular emerging phenomenon in nature: human consciousness. It is indeed the particular pattern of interactions between the systems components that gives rise to emerging dynamics which are almost impossible to be anticipated by looking only at the dynamics and characteristics the of individual elements. Neurones, in fact, posseses a fairly trivial behaviour (firing an electrical signal when the electrical voltages

at the inputs overcome a certain threshold) and certainly a neuron, taken alone and isolated from the rest of the brain, does not posses consciousness in itself. The complex dynamics of the human brain has also inspired some of the most recent advances in the domain of artificial intelligence. While the basic building blocks of neural networks have remained almost unchanged since the first *Adeline* perceptron [89], the advances in their capabilities are mostly due to novel and clever rewiring of an increasing number or artificial neurones [90]. It is worth noticing that the main difference among free-forward, associative, convolutional, and deep neural networks lies on the topologies of interconnection between the elementary blocks.

Topology plays such a crucial role that an entire line of research, *network theory*, was born with the precise goal to investigate the characteristics of the structures of relations in natural and human-made complex systems. As a result of its wide scope of applicability the interest for network science has seen an incredible growth over the last decades. The language of network science is quite universal, easily overcome disciplinary boundaries, and has taken the role of unifying framework to manage, investigate, and understand a number of diverse systems. There are precise technical reasons for which network science represents such a powerful and universally adopted tool. First, networks (or graphs) constitute a flexible and human-friendly data model which can easily catch the complexity in the data describing a number of variegate real-world systems. One of the most important technical evolutions of the Facebook platform has been the introduction of the so called *OpenGraph*, thanks to which all the data objects inside and outside the Facebook platform are described, stored and retrieved by using the simple concept of nodes and links. Also the company Google has put graph technology at the core of its

ranking algorithm since the very beginning.

Leveraging the expressive power of graph-based storage systems has rapidly become a mandatory requirement for modern companies. Alongside the giant of the Web an increasing number of smaller companies are converting their storage systems to graph technologies. Evidence of this are reported in market studies[13] and demonstrated by the proliferation of graph database solutions (e.g., NeoTechnology, OrientDB, TitanDB, FlockDB). The interest from industry has combined with those from scholars, developers and researchers, and over the last 15 years has prompted an exceptional development of software libraries which allows to perform easily computation on graphs at medium-to-large scale (FlockBD, Neo4j, Spark GraphX, GraphChi, networkx, graph-tool, igraph, SNAP, Giraph, Pragel, TitanDB). The human-friendly representation of many real-world systems as a network has made tools for graph computation attractive for data scientists as much as the wide scope of applicability of complexity and network science has made these theories popular among scholars from many disciplines.

2.5 Summary

The theoretical motivations for adopting a network approach to the description and understanding of socio-economic systems, and in particular innovation ecosystems, are even stronger than the technical ones. First above all, our society is intrinsically embedded in a network of relationships which determine social equilibria and through which information and resources are exchanged. Additionally, as transportation and communi-

[13] `https://www.forrester.com/report/Market+Overview+Graph+Databases/-/E-RES121473`

cation technologies improve and reduce in costs, the ways people interact multiply, and topologies rapidly rearrange. If, on the one hand, it is true that in the globalised world interaction is less and less influenced by geographical distances and good ideas can spread more rapidly then ever, on the other hand these processes will always be constrained by the topology of interaction. Moreover topologies co-evolve with social dynamics and it has been theorised and empirically tested [91] that particular network topologies sustain or hinder two important aspects of innovation mentioned in Section 2.3: cooperation and trust. Last but not least, comprehensive digital data on the network of interactions within innovation ecosystems is nowadays largely available and allow to quantitatively describe and understand the complex dynamics related to the creation of innovation and new knowledge. Reductionist approaches which look only at the characteristics of individual actors (e.g., scientists, entrepreneurs) will likely fail to understand and describe the complexity of innovation processes as a whole as they ignore a fundamental concept highlighted in this chapter: the locus of innovation is not in the individual but in the interaction between individuals. This argument gives a clear justification for adopting a network approach to the description of innovation ecosystems.

Chapter 3

Innovation ecosystems through the network lens

In this work we embrace a network perspective to study in detail the impact of social and professional relationships on the career of scientists and on the success of firms, in particular early-stage start-ups. In this section I will highlight the role of science and start-up businesses in innovation processes and shed light on the intimate relationship between research activities and the life in a technological start-up. I will present two case studies which will lead us to describe in details the elementary processes occurring in innovation ecosystems (e.g., exchange and recombination of ideas, knowledge transfer, collaborations patterns) and to better define the hypothesis at the root of the empirical analysis presented in Chapters 4 and 5.

Science and start-up businesses have a crucial role in our society and economy as they

are driving forces towards innovation, societal transformations and economic growth. Science has the honour and burden to push the limits of human knowledge. By doing so, scientists pave the way for unanticipated and previously unimaginable technologies, tools, and applications. Even discoveries which may seem purely theoretical at first glance can have later unforeseen impact on society. For instance, Einstein's special relativity is nowadays used to improve the precision of the satellite global positioning system (GPS). Science deals with the hardest, riskiest, but also the most long-term oriented phase of innovation. Knowledge often spills over from research centres and universities and finds its way to affect unrelated business opportunities that did not bear the cost of producing it.

While this process has been traditionally promoted with top-down approaches (e.g., departments of technology transfer looking for potential applications of the research outputs produced in their universities) in recent years bottom-up phenomena have emerged. The great propensity of academics to start their own companies has also been documented in a recent article featured on the journal *Science* [92]. The interest in "starting up" a business based on some research findings is promoted by at least two factors: (i) the inclination of funding bodies to finance applied research projects, often in partnership with industry, whose practical research outputs have more clear and direct applications into markets than theoretical ones; (ii) the low cost associated with setting up a business whose primary asset is the highly specialised knowledge of the scientists who carried the research, rather then tangible production assets. Virtually all start-ups, being in their essence knowledge-intensive activities, build on the knowledge and the results from academia. For instance, the company Google has drawn deeply in its early stages on

advances in graph theory and linear algebra, and has returned back those technologies to process extremely large amount of data (e.g., Hadoop) which are now widely used in the field of computational social science [49]. Start-ups often act as a bridge between research and society, as they are able to rapidly convert the innovation potential of some scientific advances into tangible economic growth and societal transformations.

Mainly for this reason, during the past decades, the interest in young and high-tech companies has matured exponentially among individuals, organisations, and governments. Lots of today's start-ups draw on the huge opportunities provided by the digital Era: they re-think and re-design the everyday life filling it with services, tips, recommendations, tools, advertisement, and offers. Less popular ones are focused on high-tech products for the energy, health, transport, and food sectors. For corporations and venture capitalists the term *"start-up"* refers precisely to young companies with extremely specialised and high technological profiles, equipped with business models and products capable of disrupting the current markets or creating new ones, and aiming at global and rapid scaling. Investors are tempted by the opportunity of the extremely high returns that radical new innovations may offer, while large corporations rely on various forms of external collaborations with newly established firms to outsource their innovation process and stay abreast of technological breakthroughs. For governments and individuals the term is more broadly used to refer to young entrepreneurship, including also companies which simply recombine already existing technologies into creative but relatively less scalable businesses and products. If compared with established corporations, start-ups still account for a small percentage of the richness produced by any country [93], however, from the prospective of governments, the most important element which justifies

the interest in start-ups is their role as job creators [94]. For instance, in [93] it has been shown that without start-ups there would not have been net job growth in the U.S. economy in almost every year between 1977 and 2005 [1]. This further explains why the efforts to promote and sustain young entrepreneurship and the birth of new high-tech companies are widespread in all regions of almost each modern country [94, 95].

3.1 Life in a start-up

The stories about the birth of companies such as Microsoft, Apple, Google or Facebook have led to the idea that a garage and a good idea are enough to create a billion dollar company. However, the reality in the start-up world is different from the collective imagination. As the founder of Facebook, Mark Zuckerberg, pointed out about the movie which tells the story of his company:

"The real story is a lot of hard work. If they were really making a movie (about the origins of Facebook) it would be of me sitting there coding for two hours straight."

The number of ideas turned into billion dollar companies is extremely small compared to the number of garages around the world, and start-up mortality rates are extremely high. The strategies and paths which can lead to those rare and exceptional successful companies remains elusive. In this work I put forward the hypothesis that the network of professional relationships in which the team of a company is embedded largely determines the success of a company, especially innovative ones. This hypothesis is rooted on

[1] as a reference Microsoft has been founded in 1975

the arguments presented in Section 2.3, where we have highlighted the importance that variegate knowledge and information have in the creation of innovation, and the role of network structures in securing access to them. Previous works have investigated how knowledge transfer influences firms' performance by using data from patents citations, inter-organizational collaborations, co-locations or proximity to universities as *indirect* observations of information flow [26, 96–100]. Other works have used social network analysis to study *directly* the microscopic level of interactions among individuals (e.g., inventors collaboration networks, or co-directors networks []) but have been limited to specific industries, cross-sectional observations, or small geographic areas [101, 102]. Due to the lack of data, the microscopic social dimension underlying the process of knowledge transfer, and its impact on start-up success, have remained so far largely unexplored. Before showing in Chapter 4 how large-scale digital data sets can transform our understanding of innovation ecosystems, I will describe in more details the role that professional networks play in business activities, and in particular why networking is more crucial for start-ups than for traditional business.

Here I refer to the network of a company as the collection of all relationships between the company members (e.g., founders, CEO, employes, investors, board members) and people in other organisations (other companies, governments, institutions, banks). Undoubtedly the network can have impact on the activity of any company, innovative or not. Informal relationships, in particular, constitute a valuable yet intangible asset as they facilitate the achievement of tasks, and goals. As an example, technical employees not only carry their own experience but also may provide the company with access to know-how and expertise of other organisations through their past collaborators and friends.

Sales people can leverage on their own professional and informal networks to seek for market opportunities, while human-resource departments can make use of social recruiting [2] to hire talents and facilitate their induction into the organisation. Lastly, experienced advisors can help to promptly identify and avoid unfruitful growth strategies, or speed-up the business development cycles during the quest for product/market fit, while board members in contact with institutions can provide insights on policy directions, current regulations, and negotiate on future ones by pushing towards the company interests.

To make these examples more tangible I report here two case studies concerning Uber, and the two companies Facebook and Airbnb. The evidences of the two case studies are mainly derived from the data set on start-up companies obtained from the websites Crunchbase.com and Angel.co which I will present in Chapter 4 and additional data collected from the Twitter and Linkedin platform, which has been explored only in the contexts of the case studies. First, the Facebook-Airbnb case is an example of collaboration and knowledge transfer between companies which has a clear trace in digital data available on the Web. The specific bit of knowledge we reefer to is a database technology, called PrestoDB, created by the Facebook team. The Airbnb team used PrestoDB in 2014 to create Airpal, a web-based query execution tool, which has rapidly become integral part of the Airbnb internal infrastructure with more than 1/3 of all employees issuing queries over the first year after release. [3] The more experienced Facebook team has facilitated Airbnb in the adoption and integration of the database technology, as reported in the Airbnb Web blog `http://nerds.airbnb.com/airpal/`:

[2] the process of recruiting people who are already in contact (e.g., friends of past collaborators) with current members the organisation.

[3] source: `http://nerds.airbnb.com/airpal/`

"Finally, we would be remiss if we did not mention the awesome direction that Facebook provided as the original developers of Hive and the pioneers of building UI tools to facilitate easy access to big data. We stood on the shoulders of giants to make this tool [Airpal] and we appreciate the influence and input that the data infrastructure and data tools teams at Facebook were able to provide."

These interactions between the two teams, and the associated process of knowledge transfers, have a tangible counterpart in online social networking platforms. Indeed, the data from the Twitter social network reveals 123 reciprocated following/follower connections between the Facebook and Airbnb employees, which indicates an intense exchange of information between the two groups [4]. More specifically, data from Linkedin.com reports that Mr. James Mayfield, author of the mentioned blogpost, was hired by Airbnb in 2014 after 7 years of work at Facebook. Mr. Mayfield not only has brought his expertise in his new team at Airbnb but, as we argued earlier in this section, he has also facilitated access to knowledge owned by his past collaborators at Facebook. *The role that individuals play as mediators of information and knowledge exchange* is at the core of the empirical analysis presented in Chapter 4 and 5, and it provides justification for the 1-mode projection technique used to build both the *World Wide Start-up* (*WWS*) network and the *co-authorship network* among scientists. It is also worth stressing the unprecedented resolution at which our empirical observations extend and the microscopic level of detail at which processes of knowledge transfers between companies can be observed using social-network data.

[4] We are only able to track those employees who have provided their twitter username on the platform Crunchbase.com, namely 165 Twitter users among the Facebook employees and 72 Twitter users among the Airbnb ones.

The second case study concerns Uber, the company whose smartphone App allows passengers to get in contact with private drivers. Uber has recently faced strong opposition from the main economic actor in the market of car rides: taxis. Opposition to Uber's technology has become so strong that certain countries have forbidden or are trying to forbid the company to operate in their territories. In April 2014 a Belgian court confirmed a ban on the ride-sharing App UberPOP, giving the company 21 days to close operations in Brussels threatening massive penalties [5]. Similar legal actions have been promoted in France. In 2015 Transport for London (TfL) launched a public consultation which could have resulted in a severe crackdown of Uber's activity. Uber fought back with a petition against the TfL consultation after which the transport authority decided to drop a significative number of anti-Uber proposals. In September 2015 year a California judge ruled that a lawsuit brought by Uber drivers could go forward as class-action. According to the magazine Fortune [6] it re-opened *"the biggest question about the hottest company in Unicornland: Is its business model legal? And if not, can Uber survive?"*.

Despite the strong opposition, and the severe institutional and legal obstacles that Uber's business model is facing, the company seems well prepared to fight back opposition in courts, and Uber's attractiveness for investors has not declined despite the persistent attacks from associations and institutions. The recent undisclosed Chinese investors as well as Goldman Sachs who have invested in Uber respectively $2Billion [7] and $1.6 Billion [8] have surely appreciated the strategic presence of Mr. David Pluffe in the Uber's board member of directors. Mr. Pluffe is currently referred to as the "architect" of the

[5]Source: http://www.engadget.com/2014/04/15/belgian-uber-ban-10k-fines/
[6]Source: http://fortune.com/2015/09/17/ubernomics/
[7]Source: http://www.reuters.com/article/us-uber-china-idUSKCN0UR22J20160113
[8]Source: http://techcrunch.com/2015/01/21/uber-another-1-6b/

president Obama's presidential campaign and Obama himself referred to Pluffe as the one *"who built the best political campaign in the history of the United States of America"*. It is hard to believe that the link between Uber's board and the presidential office will not affect positively the future of the company.

These two examples show how wide and strong networks of relationships can be reasonably expected to increase the chances of a company to access know-how on the latest technology, or to plan strategies to disrupt current markets and yet find the favour of governments and society. These networks, in turn, may increase the speed at which the company can executes and adapt its business plan. In today's knowledge-intensive start-ups, such networks play a fundamental role as they act as the conduits of various pools of knowledge that can be fruitfully recombined into successful innovation. As such they can even impact the survival or death of a company.

The reason why the impact of a good[9] network is so crucial for start-ups, when compared to more traditional business, is rooted in the different environmental conditions under which traditional firms and innovative start-ups operate. Often the market segments, the customers, the operations, and business models are well defined and can be clearly assessed in traditional business plans. Moreover, extremely rapid scaling or global impact are not critical requirements as in the case of start-ups. In traditional business the speed of growth can be tuned and adapted to the company's current capabilities and managers can make informed decision (e.g., on pricing) by evaluating or imitating the behaviour of competitors in the same market sector. The possibility to make reliable forecasts also mitigates uncertainty and risk and facilitates the access to debt financing.

[9] a more rigorous and quantitative definition of *good* and *bad* networks will be provided in Chapter 4

Being traditional rather then innovative does not mean being not profitable as demonstrated by the recent acquisition of the pizza chain "Franco Manca" sold for £27 Million in 2015 after 8 years of activity and 10 stores across London [10]. Business operations in a relatively stable market as the one of restaurants are radically different from the activities, the decision-making process, and management of growth in a start-up such as Uber. The pizza chain and the private-driver app have almost the same age but, in the same 8 years of activity Uber raised a total of $9 Billion and expanded in more than 400 cities around the world. Additionally, in the last few years the company was also forced to radically transform its technology, mobile application, and business models to meet legal requirements in very short times.

The book "The RainForest" by Greg Horowitt describes traditional models of business (those emerging from the Industrial Revolution) with an intriguing agricultural metaphor. A company's activity is compared to the harvesting of plantations where productivity is increased by using the latest technical tools to finely control the system (water supply, fertilisers, pesticides etc...). Once the soil quality and weather conditions are assessed the outcome is quite predictable and refined techniques can only progressively increase production, but are not likely to disrupt the production capacity or give exceptional outcomes. Buying more land will lead to an increase in production, building a faster assembly line will expand the production of cars, renting more shops will lead to an higher number of pizzas sold.

Iqbal Quadir, in his June 2015 editorial in the journal *Science*, described entrepreneurs as *"gardeners who plant innovation in the economy"*. In a similar vein, in the previously

[10] As a comparison the average amount of a start-up acquisition since 2007 was $150 Million. Source: http://techcrunch.com/2013/12/14/crunchbase-reveals-the-average-successful-startup-raises-41m-exits-at-242-9m/.

mentioned book, Greg Horowitt regards start-ups as plants in a rainforest where a variety of species continuously grow, interact and die. The great variety of species and the interaction between them favour a process that is very unlikely to happen in mass production crops: the recombination of genes and the creation of entirely new species (i.e., business ideas). Making innovative species grow in a rainforest is a task very different from harvesting crops: the condition of the soil (e.g., the market condition) are hostile and the availability of nutrients (e.g., funding) is irregular and unpredictable. Newly born species are fragile, and it is hardly predictable if the novel combination of genes will lead the plant to grow rapidly or die in the hostile condition of the rainforest.

The initial stages of a start-up business are indeed characterised by a great amount of uncertainty, unverified hypothesis, and risk. The estimation of the number of potential customers or the market volume are extremely unreliable in the early stages of truly innovative businesses, especially when the goal is to create entirely new markets from scratch. As an example, the company Facebook, originally designed to facilitate communication among college students only, has attracted the attention of more than 700 Million users across a wider range of ages and occupations and, in doing so, has created the entirely new market of social media. Another crucial risk factor is the tendency of start-up founders to overestimate the importance and pervasiveness of a certain customers' need, and consequently the potential value of the solution they offer. In reality, the actual value and the effective monetization strategies become clear only when the business face the real market and when first customers are acquired. Often, when the products hit the market, the planned strategies or even the entire business plan may need to be radically changed (as happened to the UberPOP app). The more innovative

and unanticipated the business is, the more irrelevant hypothesis and execution plans are.

To overcome the intrinsic risk of innovative endeavours several systematic methodologies have been proposed. Among them the one which has attracted the most attention is the Lean Startup Methodology, first proposed in 2008 by Eric Ries [103]. Ries' recommendation to start-ups is to adopt a combination of business-hypothesis-driven experimentation, iterative product releases, and validated learning. The central concept of the Lean Startup philosophy is to eliminate wasteful practices such as fund raising, business plans writing, or complete specifications of final product. In contrast Lean Startup promotes value-producing activities by focusing on continuous iterations of the so-called *minimum viable products* (MVPs). Building a series of MVPs rather then the final product has several advantages. Because of the mentioned uncertainty about customers and markets, it is extremely risky for a start-up to rely exclusively on the success of the first product launch, and consequently to invest all its resources on it. Instead, by introducing on the market a not finalised but cheap MVP, the company can make use of customer feedback to help further tailor the product to the specific needs emerged from the market test [67, 103]. In this way MVPs help to subject the business model to a rapid iterations of tests which allow to disprove business hypothesis and allocate resources in the most profitable way. In this sense, the process of finding the market fit matches the description of innovation as a heuristic search process (Section 2.3) in which the specific characteristics of the targeted product are only fuzzily defined and are subject to continuos revisions. For this reason, start-up teams are required to adopt a smart and creative approach with the aim to minimise effort invested in each individual

development cycle and favour instead the deployment of multiple products and market experiments. Again, this approach resemble the breath-first strategy in the search for innovative solutions mentioned in Section 2.3. Founders in particular need to understand clearly the unwarranted hypothesis at the basis of their business, and need the ability to design the right market experiments capable to test those hypothesis. Experience and feedback gained from one experiment have to be used to design the next one until a market needs are met, or new market are effectively created.

3.2 The similarities between activities in science and in start-ups

To a great extent, the condition under which start-ups operate are very similar to those of scientific research. It all starts from simple intuitions and uncertain hypotheses which need to be tested by experiments. Experimental results may be inconclusive and need to be interpreted heuristically to make new decisions about where to invest additional efforts. Decisions are often not supported by complete information and instinct plays an important role. Often the final results of research represent only a tiny fraction of all activities performed (e.g., simulations, experiments), including a large amount of intervening errors. Even though the process of ruling out wrong paths is, by definition, an important stage of research activities, the try-and-error process cannot be iterated indefinitely. Indeed, no funding bodies or investors are willing to support financially a research or a business which keep failing to meet its stated goals. If the market fit or recognition from research communities is not found within 2-3 years, the team's energy

and motivation vanish away, no matter how promising the original idea was. Successful innovative activities, either novel business models or the production of scientific articles, are those which rapidly experience the greatest number of unsuccessful paths rather then executing a fixed plan with predetermined final goals. During this frenetic try-and-error process the access to knowledge, know-how, expertise, opportunities is crucial because it reduces the chances to take unfruitful path. Scientific collaborators or members of the same research group may share their experience and use their intuition to guide projects towards more fruitful and promising activities. This is analogous to the benefits that a start-up can gain from the experience of advisors and mentors in the decision-making process and the phase of market experiments.

3.3 Life in academia

Just as knowledge flows could be mapped in the context of start-up firms, in a similar way data enable us to track exchange of knowledge among individuals in the domain of science. In their daily academic lives, scientists typically exchange expertise, discoveries, techniques, and knowledge of the literature. These exchanges may occur in different contexts such as conferences, workshops, meetings, or they may even occur as scientists read one another's articles or manuscripts. Thus, as stated in Section 2.1 and 2.3, all these exchanges have a strong social connotation.

A great portion of the whole academic life (scientists' research interests, prestige, collaboration patterns) leave clear traces in digital data about scientific publishing. Since scientific production has increased at a fast pace over the last few centuries, and since

nobody can master or contribute to all topics at the same time, the scientific enterprise has rapidly converged to a structure organised into disciplinary silos which is reflected mainly by the subdivision of modern universities into academic departments. Having defined boundaries makes it easier and more effective for scientists to draw deeply on the knowledge accumulated over centuries, and collaborate with other researchers in their chosen disciplinary silo. The life of scientists in modern academia is remarkably different from that of, for instance, ancient Greek philosophers or great scientific minds such as Leonardo Da Vinci, who were able to master the breadth and depth of the whole disciplinary landscape. Even within a specific discipline, the amount of knowledge in the various subfields is so vast that scientists are forced to specialise into narrower and narrower research areas in order to able to provide their contribution. For instance, Physics is broadly divided into electromagnetism, optics, acoustics, dynamics of fluids, condensed matter, nuclear physics, elementary particles, atomic and molecular physics, geophysics, astronomy, and astrophysics. However, grasping in full depth all knowledge of a subfield is a challenge even for senior researchers.

Communication across the boundaries of disciplines and sub-fields can provide significant advantage as it offers more opportunities to recombine ideas, stimulate creativity, and ultimately produce new knowledge and innovations. The comprehensive nature of interdisciplinary research (IR) is also expected to enable scientists to solve complex global issues which cannot be addressed by those who only embrace a specialised disciplinary perspective. The journal Nature, in a recent special issue [104], has regarded interdisciplinary researchers as the super-heros of science and considered IR essential *"to solve the grand greatest challenges facing society: energy, water, climate, food, health"*. There

have been all along history periodic efforts to promote collaboration and communication across academic disciplines. Jacobs and Frickel [105] date the push towards interdisciplinary research back to the work of the Social Science Research Council of the U.S. in the 1920s, and to the Rockefeller Foundation in the 1930s. Evidence of the most recent push is found in the increasing number of interdisciplinary training programs, research consortia, scholarships and funding opportunities provided by national research agencies.

The modern trend of interdisciplinary research has also generated intense debate and several criticisms. Because of the absence of strong theoretical arguments and empirical evidence, skepticisms have emerged on the assumptions advanced by advocates of interdisciplinarity, and on the claimed superiority of interdisciplinarity over disciplinary knowledge [105, 106] . Indeed, several costs and disadvantages may also arise from abandoning specialisation in favour of a more inclusive and integrative approach to research. Large and heterogeneous research teams may experience coordination and communication obstacles that can outweigh the advantages of the diversity in backgrounds and skills. If, on the one hand, diversity provides more opportunities for novel recombinations of ideas, on the other hand it carries the extra costs of dealing with the different research approaches, languages, techniques and style of thoughts, and past literature of each discipline. The mentioned obstacles have often been used to justify a series of empirical evidence supporting the idea that scientific achievements are obtained at intermediate levels of interdisciplinarity and that highly interdisciplinary articles that mix together disparate knowledge pools usually produce a lower scientific impact than articles concentrated on fewer topics [18]. The most recent literature [18, 107, 108] has mainly focused on providing measure of the degree of interdisciplinarity of scientific

activities (e.g., production of articles and research projects). In these studies the concept of interdisciplinarity has been operationalised in various and often not comparable ways, mainly drawing on the list of references to construct interdisciplinarity scores. However, the particular operationalisation of interdisciplinarity which draws on the reference list *disregards the actual breath of knowledge of individual authors and the intensity of the knowledge exchange through scientific collaborations (e.g., through the co-authorship of an article).*

Even though numerous works have explored the comparative advantages of interdisciplinary research in general, only a few works have focused attention on individuals and studied interdisciplinarity across career trajectories [109–111]. The current literature does not still clearly establish whether pursuing interdisciplinarity provides advantages to an individual researcher. In particular there is no clear evidence on whether or not it is more effective for researchers to specialise in narrow fields than to have broad scientific interests. Additionally, as I stated in Section 2.3, innovation and novelty result from the recombination of ideas draw from the network in which authors are embedded. Yet, few works have investigated the role of the network in fostering interdisciplinarity and its relation to authors' scientific performances.

Given the institutional pressure for interdisciplinary projects on the one hand, and the risks associated to them on the other, it has become crucial for scientists to know which are the most effective mechanisms, career paths and research strategies that nurture and sustain scientific success in modern academia. Do researchers need to overcome disciplinary boundaries and abandon specialisation in favour of more inclusive perspectives to amplify their scientific impact and recognition? To which extent does this attempt pay

back the higher costs associated with interdisciplinary research? Can authors preserve their specialised approaches while leveraging on heterogeneous expertise of collaborators in order to pursuit IR and avoid knowledge overloads? How is a scientist's personal expertise influenced by the knowledge of the those with whom the scientist collaborates? And which collaboration patterns are more effective in sustaining scientific performance? I try to answer these questions in Chapter 5.

3.4 Summary

Science and start-ups have a crucial role in our society and economy as they are driving forces towards innovation, societal transformations and economic growth. The process through which scientists and entrepreneurs produce innovation are intimately related. Their common denominators are: uncertainty of outcomes, trial-and-error processes, access to complex, multi-faceted and often tacit knowledge, and the role of advisors, mentors, and scientific collaborators in speeding up experimental phases. In this work I embrace a network perspective to study the impact that social and professional networks have on the performance of scientists and start-ups. The overall goal is to use digital data to design methodologies to guide and support innovation processes systematically, and reduce the intrinsic risk associated with innovation. To conduct this empirical investigation I have collected, cleaned, and analysed some of the largest data sets publicly available about innovative activities. In particular I have focused my attention on the data available from the websites of the following organisations: Thomson Reuters' Web of Science (WOS), American Physical Society (APS), Cruchbase.com, Angelist.co. The

next two Chapters present the empirical results based on these data sets.

Chapter 4

Empirical investigation I:

Predicting the success of start-ups

4.1 The Crunchbase data set

Data were collected from the Crunchbase.com Web API and are updated to December 2015. The data on the Crunchbase website are manually curated by several contributors affiliated with the Crunchbase platform (e.g., incubators, venture funds, individuals) and are enriched by automatic crawlers which scrape the Web, on a daily basis, searching for news about initial public offers (IPOs), acquisitions, and funding rounds. To date Crunchbase is widely considered the world's most comprehensive open data set about start-up companies. For each organisation in the Crunchbase data I extracted all the people included in the team (e.g., founders, advisors, board member, employees), and additional information such as foundation date, location of the firm headquarters,

founding rounds, acquisitions, and IPOs. Organisations and people are uniquely identified by alphanumeric IDs. Organisations belong to four categories: companies, investors, schools, groups. Among the schools I counted 383 universities, including top tier institution such as Stanford University, the Massachusetts Institute of Technology (MIT) and many others. Hence, in addition to people's business activity, the data track also information about their education paths, and consequently the possibility to draw knowledge from academia. All organisations have been considered in the analysis, but only those belonging to the category *"companies"* have been included in the ranking method illustrated in Section 4.3. All data are time-stamped and an accurate reconstruction of historical events is made possible by the use of *trust-codes*, i.e., numerical codes provided by Crunchbase to indicate the reliability of a certain timestamp. The timestamps indicate the dates of foundation, funding rounds, acquisitions, and IPOs. Raw data were stored on a Neo4j graph database instance from which I constructed a bipartite time-varying graph with $41,830$ nodes representing firms distributed across 117 countries around the globe, $36,278$ nodes representing people, and $284,460$ links between people and companies. The graph is time-varying because each node and each link have a time associated, representing the time a person held a professional role in a company. Notice that in the construction of the time-varying graph I retained only the timestamps whose trust-code guarantees the reliability of year and month.

Additionally, I cleaned the data by solving and removing inconsistencies such as an employee's role starting at a date prior to the company's foundation. In these cases I retained the most reliable information according to the trust-code value. Inconsistencies were removed by adopting a strong self-penalising cleaning strategy: I did not make any

assumption on timestamps nor did I try to infer them and I did not retain in the graph links whose timestamp could not be determined in a reliable way. This data cleaning approach strengthens the validity of the results in Section 4.3 because it does not enable companies to gain positions in the closeness centrality rank score thanks to connections which were been forged at later time but were incorrectly or partially reported in the data. In this way I avoid biases that can artificially inflate the success rate of the method presented in Section 4.3. The cleaned data will be made available for public use on `http://maths.qmul.ac.uk/~mbonaventura`.

Figure 4.1 illustrates the coverage of the Crunchbase data around the world. The image shows the density of companies in each city of the world (close cities are aggregated for layout purposes). Beyond the unsurprising predominance of the U.S. ecosystems, it is possible to notice an higher density around the capitals of Europe, led by Germany and the United Kingdom, as well as the Israeli innovation cluster. Few denser clusters can be found also in Australia, East Asia and South America.

Figure 4.2 shows the Airbnb neighbourhood in the two-mode graph, and illustrates the microscopic resolution at which the Crunchbase data enables one to track the potential flow of resources (e.g., knowledge and information) between companies. In 2013 Airbnb hired Mr. Thomas Arend (highlighted in the red square), who had previously acted as a senior product manager in Google, as an international product leader in Twitter, and as a product manager in Mozilla. The professional network thus reveals the potential flow of knowledge between Airbnb and the three other companies in which Mr. Arend had played a role.

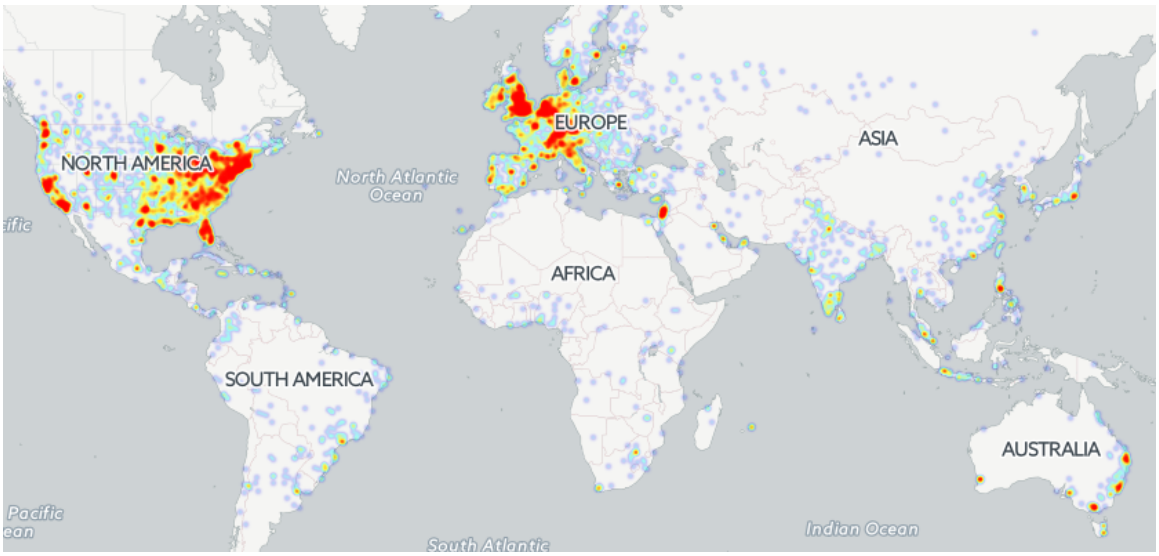

Figure 4.1: Distribution of startup companies around the world in 2000. Notice the higher density around the capitals of Europe, as well as the Israeli innovation cluster.

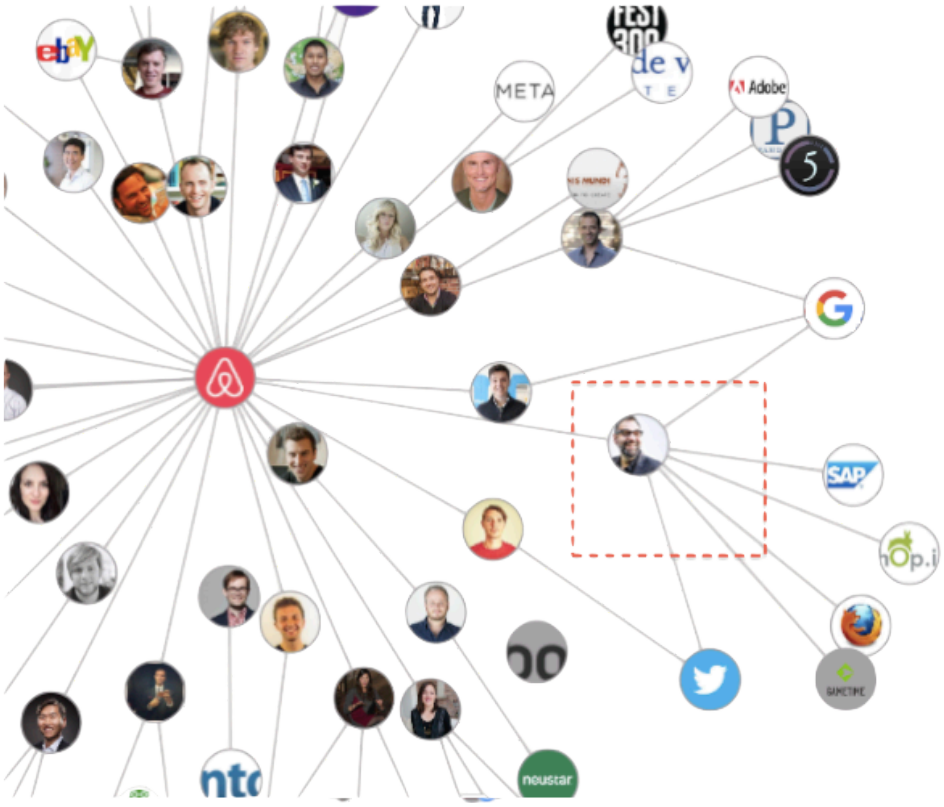

Figure 4.2: The Airbnb neighbourhood of the two-mode graph in 2013. Mr. Thomas Arend, highlighted in the red square, was hired in 2013 and secured the potential for knowledge flow from Google, Twitter, and Mozilla where he previously acted as senior product manager.

I have conducted two main empirical studies which make use of temporal data, geographic data and the underlying graph of relationships among people and companies: the *Success prediction method* presented in Section 4.3 and the *Start-up fingerprints of cities* presented in Section **??**.

I have also integrated the Crunchbase.com data sets with data from the platform Angel.co collected through their Web API. Since the two websites have stipulated a partnership which allows them to import their respective data, the majority of the information in the Angel.co platform are actually just a copy of the data included in the Crunchbase one. However, while on Crunchbase the various professional roles are identified by free-text strings, and consequently may include misspellings, in the Angel data set the roles are conveniently collected in 6 unique categories: founder, employee, advisor, investor, mentor, and attorney. These unique identifiers are extremely convenient as they enable us to carry out the analysis present in Section **??**. Crunchbase data contains more reliable information about funding rounds, IPOs and acquisitions, and more reliable timestamps which are crucial in the longitudinal analysis of the success prediction method presented in Section 4.3.

4.2 The World Wide Start-up (WWS) network

I project the bipartite time-varying graph into a one-mode graph in which two companies are connected when they share at least one individual that plays or has played a professional role in both companies. Such a graph comprises $\mathcal{N} = 41,830$ companies and $\mathcal{K} = 135,099$ links among them, and has been named the *World Wide Start-up* (*WWS*)

network. The projected graph is time-varying like the original bipartite graph: a link between any two companies is forged as soon as one individual with a professional role in one company takes on a role in the other company. Data cover an observation period ranging from 1990 to 2015 at a monthly resolution. In this period, various communities of start-ups around the globe have joined together to form the largest connected component which in 2015 includes about 80% of the nodes of the WWS network. Moreover the merging of various components over time has reduced the *"degree of separation"* between any two companies in the WWS network to an average of 4.74. Figure 4.3(a) highlights the countries in which start-ups have joined, over time, the largest connected component of the WWS network. In Figure 4.3(b-c) I report, as a function of time, the number of nodes and links in the WWS network and the fraction of nodes belonging to the largest component of the graph. In particular Figure 4.3(b) indicates a steady exponential growth in the number of companies over the last 25 years.

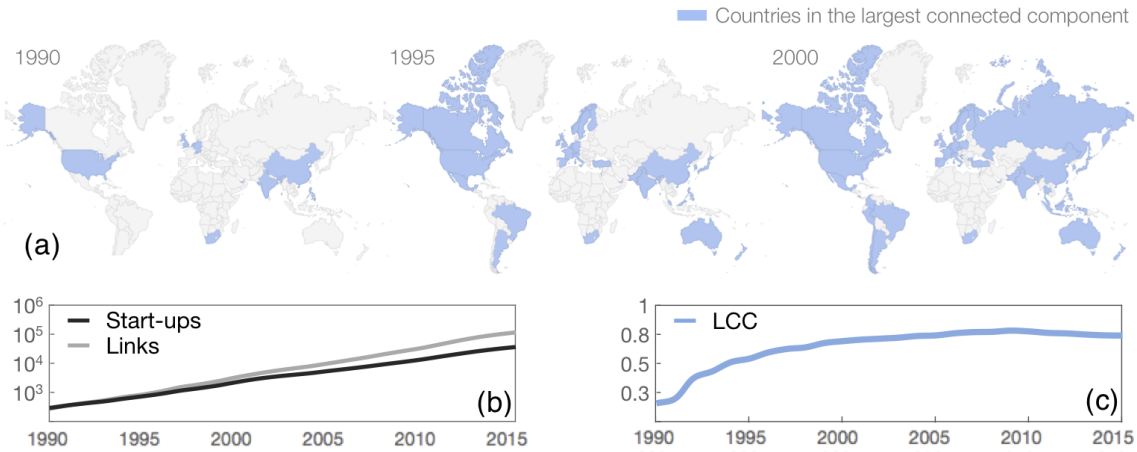

Figure 4.3: (a) Countries that, over time, joined the largest connected component (LCC) of the worldwide start-up (WWS) network are highlighted in blue; (b) evolution of number of firms and links in the WWS network; (c) fraction of nodes in the LCC over time.

4.3 Predictions of success

4.3.1 Methods and measures

As new links are forged over time, the distance from a certain company to all other firms in the WWS network reduces, which in turn enables the company to gain new knowledge and tap business opportunities beyond its immediate local neighbourhood. It is reasonable to expect that companies with a central position in the WWS network can benefit from greater knowledge transfer and easier access to resources and opportunities. The concept of centrality and network measures were introduced in the context of social network analysis, and more recently have been applied to various other fields [43, 112]. To capture potential exposure to knowledge, I have computed, for each month, the closeness centrality of each node in the WWS network. The closeness centrality $C_i(t)$ of a node i, $i = 1, 2, \ldots, \mathcal{N}(t)$ quantifies the importance of a node in the graph by measuring its mean distance from all other nodes and is defined as:

$$C_i(t) = \frac{\mathcal{N}(t) - 1}{\sum_j d_{ij}(t)}, \tag{4.1}$$

where $\mathcal{N}(t)$ is the number of nodes in the graph at time t, while $d_{ij}(t)$ is the graph distance between the two nodes i and j, measured as the number of links in a shortest path between the two nodes. To account for multiple disconnected components I have used a generalisation of the original closeness centrality as proposed in [113]. In each month of the observation period, I ranked companies according to their values of closeness centrality (i.e., top nodes are firms with the highest closeness). I refer to the ranked

list which include all companies as the *full list*. Figure 4.4 shows an example of the

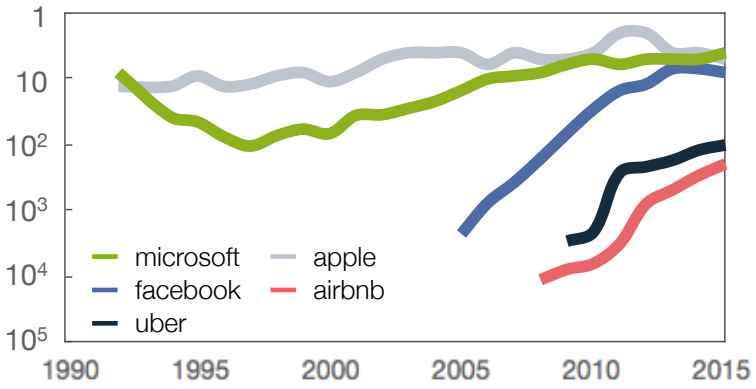

Figure 4.4: Evolution of closeness centrality rank of five popular firms.

large variety of observed trajectories as companies moved towards higher or lower ranks, i.e., they obtained a larger or smaller proximity to all other companies in the network. For example, Apple has always been in the top 10 firms over the entire period, while Microsoft exhibited an initial decline followed by a constant rise, moving the company towards the central region of the network. The trajectories of younger start-ups, such as Facebook, Airbnb, and Uber, are instead characterised by an abrupt and swift move to the highest positions of the ranking soon after their foundation, possibly as a result of the boost in activity that has characterised the venture capital industry in recent years. For instance, the sudden jump at the beginning of the Uber trajectory is due to the first $1M$ U.S. dollars investment round joined by 17 distinct investors.

To investigate the interplay between the position of a company in the WWS network and its long-term success I have used additional data on funding rounds, acquisitions, and initial public offerings (IPOs) collected through the Crunchbase Web API. This data have been coupled with the closeness centrality ranking to construct the prediction method described in what follows.

4.3.2 Predictions

For each month t, I constructed an ordered list of $N(t)$ firms, ranked by closeness, that can be classified as *"open deals"* for investors, i.e., they have not yet received funding, have not yet been acquired, and have not yet been listed in the stock exchange market. As an example, the company WhatsApp, which in June 2009 had not received any investment and ranked 1060th in the full list, occupied the 15th position in the open-deals list in the same month. Notice that, by assessing a firms network position prior to any funding event or IPO, the analysis is not subject to possible biases arising from the effects that the capital market might have upon the firms expected performance. Figure 4.5 shows an illustrative example of the construction of the open deal lists (the company names do not reflect the real data). The figure illustrates the two cases in which a company is removed from the open-deal list: (i) the company receives funding, and (ii) it is older than 2 years. I then identify which companies, within a time window $\Delta t = 7$ years starting at month t, succeed in securing at least one of the following *positive outcomes*: (i) the company makes an acquisition; (ii) it is acquired by another company; or (iii) it undergoes an IPO. Company success is hence regarded as binary variable equal to 1 if the company has achieved a positive outcome or 0 otherwise. While the fact that a company is acquired or undergoes an IPO signals a potential and measurable economic gain for the company shareholders, I have used the acquisition of other organisations as indirect measure of the company's economic success. Indeed, in the absence of data on revenues, the acquisition of other organisations is the only event which can signal a solid financial status and growth. More importantly, without this third indirect measure of company success, we would have considered as unsuccessful companies with skyrocketing

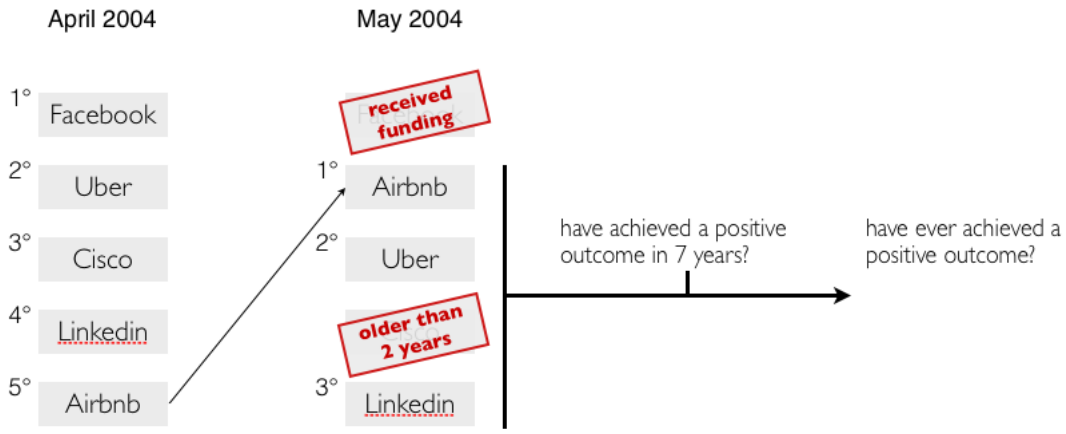

Figure 4.5: Illustrative example of the construction of the open-deals lists and the prediction methods (company names do not reflect the real data). In May 2004 Facebook has received funding while Cisco is older then 2 years. Both companies are then removed from the open deal list and the ranking is updated. In each month I check whether or not a company has achieved a positive outcome within 7 years or, alternatively, ever in the future.

revenues such as Airbnb, Uber, and Deliveroo.

To asses the accuracy of the method in identifying successful companies, I check how many of the Top 20 companies ($n = 20$) in the closeness-based ranking of open-deals have obtained a positive outcome. In particular I have computed the success rate $S(t)$ defined as the ratio $m(t)/n$, where $m(t)$ is the number of firms with a positive outcome included in the Top 20 ($n = 20$) of the open-deal list of month t. If the open-deal lists were randomly ordered, the expected number $m^{\text{rand}}(t)$ of successful companies in the Top 20 would have been given by the expected value of the hypergeometric distribution $\text{H}(N(t), M(t), n)$, where $N(t)$ is the total number of companies in the open deal list, $M(t)$ is the total number of companies which have achieved a positive outcome, and $n = 20$ is the length of the Top list considered. In particular, in the case of random ordering, the expected success rate $S^{rand}(t)$ is given by $S^{rand}(t) = M(t)/N(t)$. The statistical significance of the success

rate is assessed by computing the hypergeometric p-values, which give the probability of obtaining, by chance, a success rate greater than or equal to the one obtained empirically. Such probability can be written as $p(t) = \sum_{k=S(t)}^{M(t)} \mathcal{H}(N(t), M(t), n, k)$, where $\mathcal{H}$ is the probability mass function of the hypergeometric distribution.

4.3.3 Empirical findings

Figure 4.6(a) compares the actual success rate $S(t)$ (blue curve) of the prediction method, with the one expected by chance $S^{\text{rand}}(t)$ (black curve). The p-values in the top panel of Figure 4.6(a) measure the probability of obtaining, by chance, a success rate larger than $S(t)$, with low values of p (highlighted regions) indicating the time periods where the prediction is statistically significant. From mid-2001 to mid-2004, the success rate (blue curve) is remarkably larger than the one based on random expectations (black curve), and the p-value is always smaller than 0.01. The success rate exhibits also an exceptional peak of $\sim$ 50% in June 2003 (p-value = 0.0001). In this month the ten companies in the Top 20 of the open-deals list are: Mailfrontier, Proofpoint, Riverbed Technology, Bluelane Technologies, Xfire, Loyalty Matrix, Verdisoft, Instore, Dupont Photonics, Istante Software. From 2004 to 2007, the blue curve decreases, reaching a local minimum at a time when a global financial crisis was triggered by the US housing bubble. In this period (as well as during the collapse of the dot-com bubble in 1999-2001), even though the success rate still exceeds random expectations, the high p-values indicate that the discrepancy between $S(t)$ and $S^{\text{rand}}(t)$ are not statistically significant. Finally, after mid-2007, the performance of the prediction increases, and it stabilises around 35% (p-value = 0.01).

Figure 4.6(b) shows the success rate $\widetilde{S}_I$ aggregated over all observed periods in the following manner. The overall success rate $\widetilde{S}_I$ takes into account the total number of positive entries in the Top of all open-deal lists, regardless of the specific companies which occupy those positions. In this way $\widetilde{S}_I$ provides a measure of the overall goodness of the ranking across months, but it does not provide information about the number of unique companies correctly or wrongly identified as successful. As an example of the computation of $\widetilde{S}_I$ let us consider the period starting in January 2000 and ending in December 2007, and the Top 20 (bottom-left plot in Figure 4.6). Such a period includes $\delta = 96$ months. The overall success rate $\widetilde{S}_I$ is defined as:

$$\widetilde{S}_I = \frac{\widetilde{m}_I}{\widetilde{n}_I}$$

where $\widetilde{n}_I = 20 * \delta$ is the total number of entries in the Top 20 across the δ months, while $\widetilde{m}_I = \sum_t m(t)$, where the sum runs over all the months in the considered period. If we randomly shuffle the entries within a month, i.e., we shuffle each open-deal list independently, the expected total number of successful companies within all the Top 20s is given by:

$$\widetilde{m}_I^{\text{rand}} = \sum_t m^{\text{rand}}(t)$$

and the corresponding variance is given by the sum of the variance in each month

$$Var(\widetilde{m}_I^{\text{rand}}) = \sum_t Var(m^{\text{rand}}(t)).$$

The expected overall success rate in the case of random ordering is then given by

$$\widetilde{S}_I^{\text{rand}} = \frac{\widetilde{m}_I^{\text{rand}}}{\widetilde{n}_I}$$

and its standard deviation σ_I is:

$$\sigma_I = \frac{\sqrt{Var(\widetilde{m}_I^{\text{rand}})}}{\widetilde{n}_I}.$$

Figure 4.6(b) reports the overall success rate empirically found $\widetilde{S}_I$ (blue bars), the overall success rate $\widetilde{S}_I^{\text{rand}}$ expected by chance (black dots), and its standard deviation (black error bars) for different numbers of recommended companies (i.e., Top 20,50,100).

4.3.4 Robustness

I checked the robustness of the prediction method by replicating the analysis based on the $n = 20$, $n = 50$, and $n = 100$ firms with the largest closeness centrality, and two different time windows, namely $\Delta t = 6$ and $\Delta t = 8$ years. The panels in Figure 4.8 report the evolution of the success rate and the p-values over time for $\Delta t = 6, 7, 8$ and $n = 20, 50$. The overall trend and the minima of the success rate during the periods of market instabilities are confirmed for all time windows and values of n.

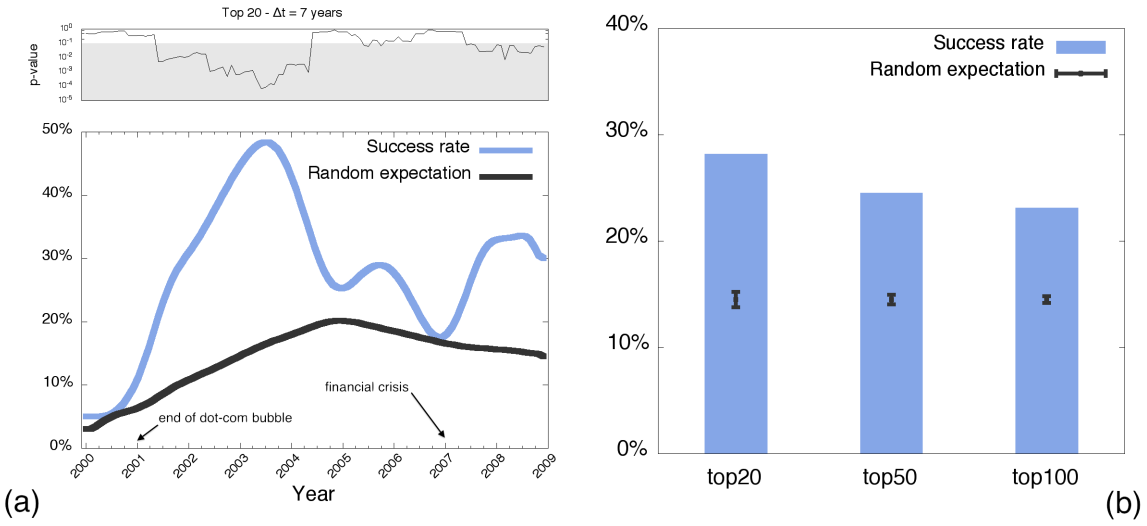

Figure 4.6: (a) The success rate $S(t)$ of the prediction method (blue curve) is compared to the expected success rate $S^{\text{rand}}(t)$ of a random selection of the recommended companies (black curve). The success rate reaches an exceptional peak of about 50% around June 2003, and reduces significantly in correspondence of periods of financial instability (dot-com and housing bubble). The statistical significance of the discrepancy between $S(t)$ and $S^{\text{rand}}(t)$ is quantified by the computation of the associated p-values, shown in the top charts of each panel. (b) The success rate $\widetilde{S}_I$ aggregated over the entire observation period. Notice the discrepancy between the actual success rate of the prediction method and the random expectation. Increasing the length of the Top list considered (n from 20 to 100) the aggregate success rate decreases indicating that the highest positions in the list are characterised by an overabundance of successful companies compared to the number expected in case of random ordering of the list.

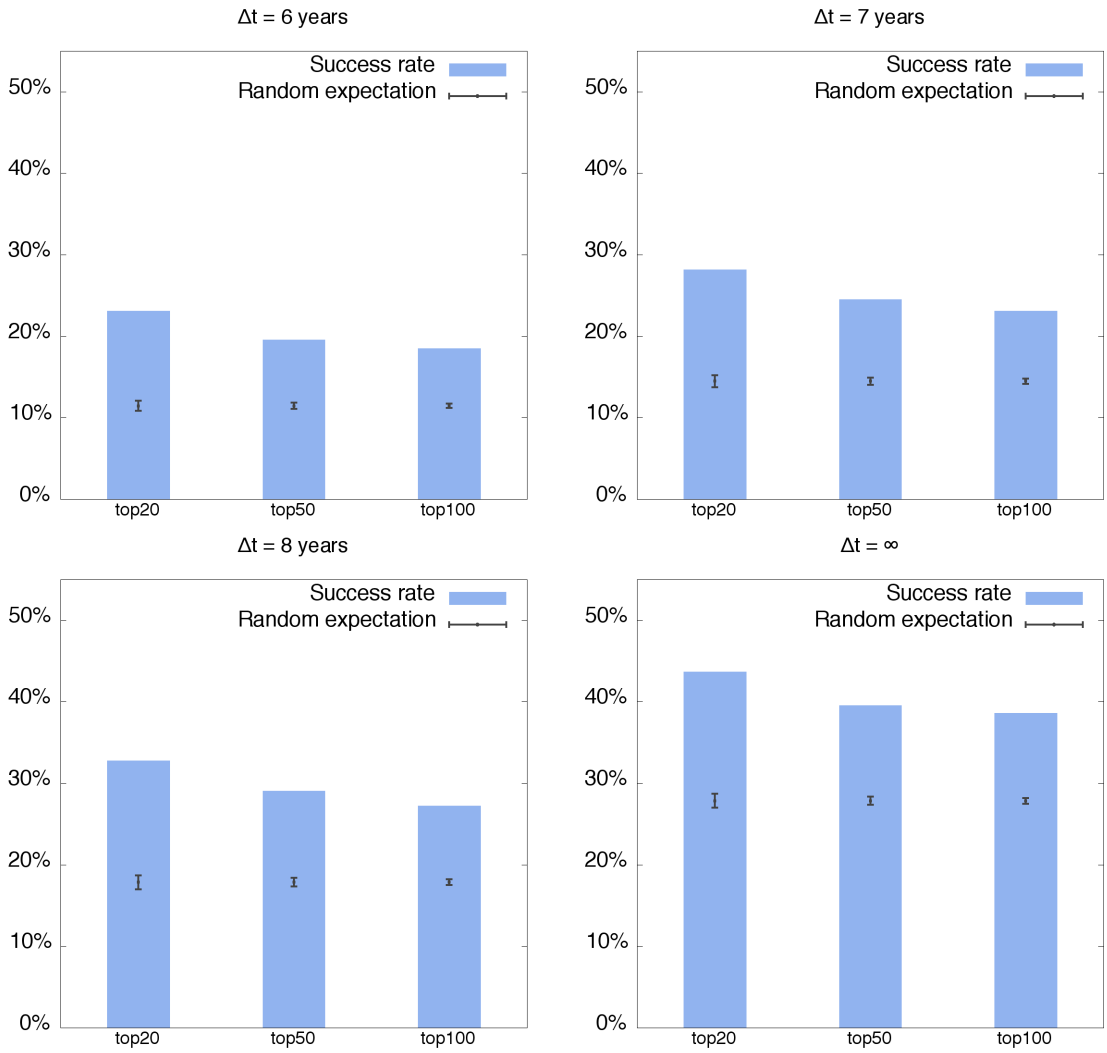

Figure 4.7: For robustness check I replicated the analysis of fig 4.6(a) for different time windows, $\Delta t = 6, 7, 8$ years. The bottom-right panel shows an additional analysis in which I remove the restriction on the maximum time in which a company is allowed to achieve a positive outcome, i.e., I set $\Delta t = \infty$.

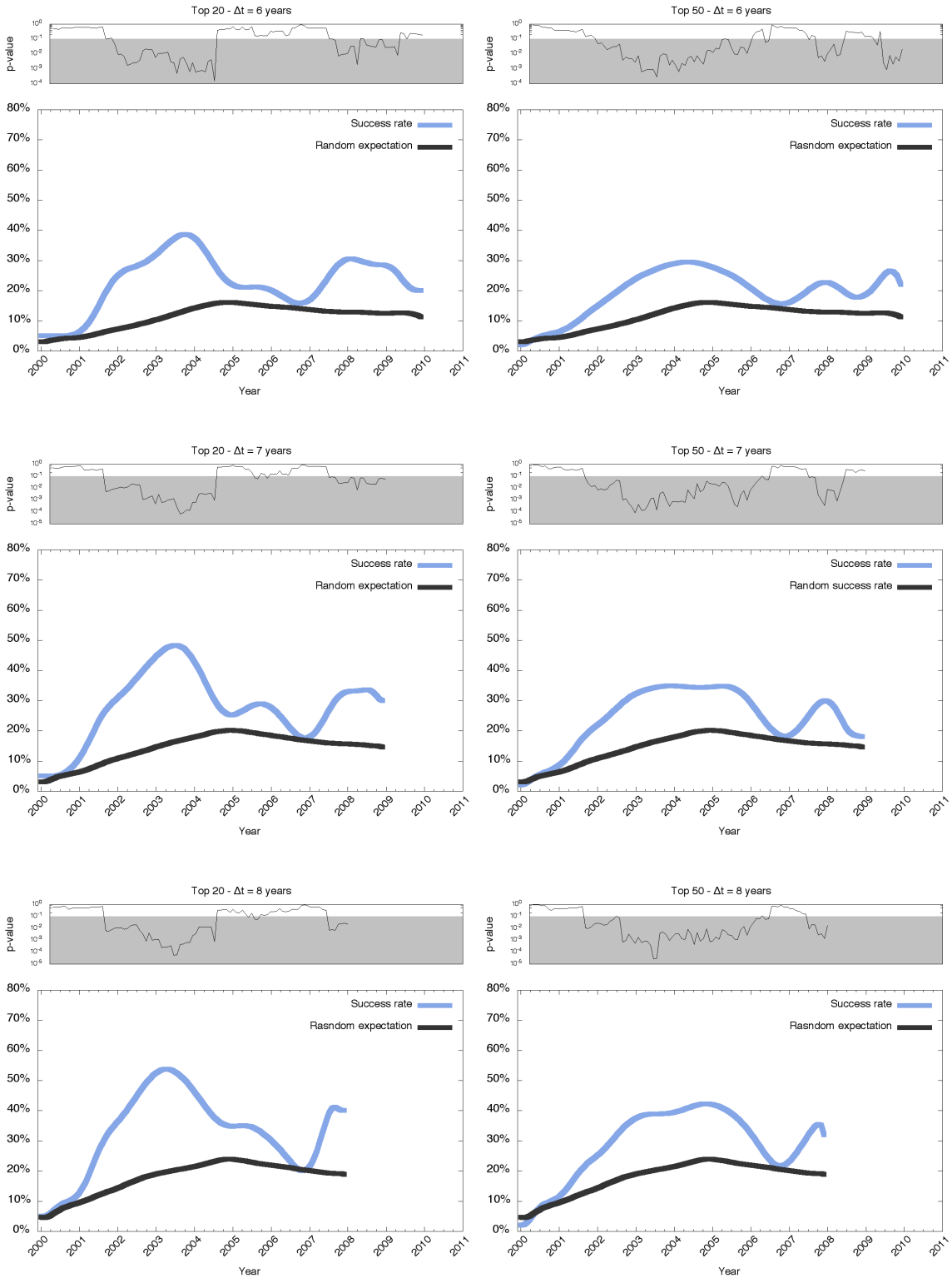

Figure 4.8: For robustness check I replicated the analysis of fig 4.6(a) for different lengths of the recommendations list, namely $n = 20, 50$, and different time windows, $\Delta t = 6, 7, 8$ years. The trend of the blue curve is consistent across all the analyses and the local maxima and minima are not influenced by the particular choice of the parameters.

4.4 Discussion

In this chapter I have shown how today's digital data allows to investigate and characterise innovation processes at a global scale. I have presented an integrated approach to collecting, analysing, and makings sense of data about start-up ecosystems. Based on the hypotheses and methodological framework presented in Section 2.4 I have adopted a network perspective and used the data to construct the WorldWide Startup (WWS) network. Mapping the potential knowledge and information flow in terms of a network has proven to be an extremely powerful approach. First, the network perspective has allowed me to reveal differences and similarities between cities around the world and to distinguish the various patterns of activity of different ecosystems. This study provides ecosystem managers with a quantitative and synthetic way to monitor their local startup communities and benchmark them against well-know successful clusters such as the Silicon Valley. More importantly, I have shown a direct relation between the position of a company in the WWS network and its long-term performance. In particular I have shown that companies which occupy central positions in the global network, i.e., have higher closeness centrality scores, have higher probability to be successful than companies with less central positions. The analysis and methodologies proposed have various policy implication and significant societal and economic impact. The long-term vision of these studies is to improve the way institutions timely gather and make sense of data on innovation ecosystems, and to provide science-based methodological frameworks to optimise the investments of governments and individuals into innovative activities. In 2014 the European Commission made a first step in this direction by commissioning the platform Startuphubs.eu, whose scope remains limited to aggregating demographics

and statistics on a few EU cities. My work contributes in this direction by integrating raw demographic data with the important layer of professional interactions, and by proposing a multidisciplinary perspective which combines network theory, social science, data science, and the science of success. The approaches and methodologies developed in my research can help to understand, mitigate, and better manage the risk associated with the investment in early-stage innovative activities. This can enable the managers of investment funds to: (i) raise more capital and reduce the financial gap that divides European countries from other, more risk-prone, countries such as the US; (ii) increase the number and the amount of investments in early-stage companies that will shape our technological future over the next years; (iii) increase the return on investment and capitalisation of highly risky start-up firms.

Chapter 5

Empirical investigation II: The advantages of interdisciplinarity in modern science

The dichotomy between specialisation and interdisciplinarity in modern science is a topic of controversial and ongoing discussion. On the one hand, modern scientists are pushed towards specialisation by the difficulty in mastering the huge amount of knowledge accumulated over time and by the institutional consolidation of scientific domains. On the other hand it is widely recognised that the recombination of knowledge from different disciplines can allow to tackle complex problems [104], boost the generation of novel ideas, and enhance scientists' career. However, abandoning specialisation does not come without risks. As already discussed in section 3.3 the difficulties in combining research approaches, languages, techniques, and styles of thought of different disciplines can out-

weigh the advantages of backgrounds and skills diversity. Which are, from the prospective of individual researchers, the most effective and successful ways to overcome disciplinary boundaries? The goal of the analysis presented in this chapter is to understand the impact of interdisciplinarity on scientific success, and in particular to assess under which condition performing interdisciplinary research has a positive effect on the number of citations accumulated by an author or if, instead, scientists specialised on a specific subject have a higher probability of getting more citations than interdisciplinary ones. I try to shed light on this matter by looking at the largest publicly available data sets on scientific collaborations. I analyse the co-authorship network, citations and the information on sub-field classification extracted from the bibliographic data set of all journals included in the American Physical Society (APS) and Web of Science (WOS). To this end I also propose two novel measures of interdisciplinarity: the background and social interdisciplinarity.

5.1 The APS and WOS datasets

The analysis draws on two databases on scientific publishing: the American Physical Society (APS), and the Web of Science (WOS). Table 5-A reports summary statistics on the two data sets.

Data set	**Authors**	**Articles**	**Years**	**Research categories**
APS	136,871	380,913	1980 - 2014	10 / 1,154 (broad / fine-grained)
WOS	602,299	1,125,729	1945 - 2014	50

Table 5-A: Summary statistics on the APS and WOS data sets.

APS data set. The APS data set includes bibliographic information on all the articles

appeared in the scientific journals published by the American Physical Society between 1893 and 2014 [1]. For each article, I extracted the title, the abstract, the research subject(s), the name(s) and affiliation(s) of the author(s), the publication time, and the number of citations received. Each article published after 1980 is associated with at least one and up to four codes included in the Physics and Astronomy Classification Scheme (PACS). PACS codes identify the sub-field(s) of physics to which each article has contributed. The structure of the PACS codes is hierarchical, and consists of 10 top-level categories which split up into two further levels. For example, the PACS code "87.14.ep" identifies articles in the sub-category "Membrane proteins", which belongs to the category "87.14 - Biomolecules: types", which in turn belongs to the sub-field "87 - Interdisciplinary physics – Biological and medical physics". PACS codes are chosen by authors based on the list provided on the APS web site. Notice, however, that the appropriateness of the choice of PACS codes is typically assessed by reviewers and the editorial office during the revision process. In the analysis I considered only the first two hierarchical levels of PACS codes, thus obtaining $M = 1,154$ distinct PACS codes. I restricted the analysis only to articles associated with PACS codes, i.e., published after 1980, and to authors whose careers started after 1980. I also filtered out all the articles authored by more than 10 co-authors, typically resulting from large-scale experiments in high-energy physics. Authors are identified by surname, and first and second names (in some cases only the initials of first and second names were available). I employed three different name disambiguation strategies that, respectively, take into account: (i) the entire surname and name initial(s), (ii) the entire surname and the entire name, (ii) the entire surname, the entire name or its initial, the affiliations, and collaboration

[1] `https://publish.aps.org/datasets`

and citation networks. The analysis presented in this chapter was based on the third disambiguation method. The resulting data set includes $N_P = 380.913$ articles and $N_A = 136.871$ authors (i.e., 35% of the total unfiltered number of authors). The details of the various disambiguation methods here used, and the robustness of the results obtained through such methods are discussed in Appendix C.2.

WOS data set. From the WOS webpage[2] I have manually collected bibliographic information on the articles belonging to the 50 research categories listed in Table 5-B. For each research category, I identified the top 5 scientific journals with the highest impact factor. I then downloaded metadata on articles published in these journals. For each journal, at least one of the following conditions was satisfied: (i) all articles published in the journal were downloaded, or (ii) at least $20,000$ articles published in the journal were downloaded, or (iii) at least all articles published in the journal over the last 20 years were downloaded. The final data set includes $N_P = 1.125.729$ articles published between 1945 and 2014. For each article, the following information were available: title, publication time, full name(s) of the author(s), and the total number of citations received up to March 2015. Additionally, each article is associated with one of the 50 research categories listed in Table 5-B. Because additional information, such as records on institutional affiliation, could not be retrieved from the WOS, I were able to employ only an initial-based name disambiguation strategy [114], thus obtaining $N_A = 1.532.673$ unique authors (i.e, 39% of the total unfiltered number of authors).

Since research categories in the WOS data set play the same role as PACS codes in the APS data set, all measures introduced in Section 5.2 can be easily applied to the

[2] `http://apps.webofknowledge.com`

Biology	biology biochemistry molecular biology biotechnology applied microbiology biophysics cell biology cell tissue engineering biochemical methods
Chemistry	chemistry analytical chemistry applied chemistry inorganic nuclear chemistry medicinal chemistry multidisciplinary chemistry organic
Computer science	artificial intelligence cybernetics hardware architecture information systems interdisciplinary applications software engineering theory methods
Mathematics	mathematical computational biology mathematics mathematics applied interdisciplinary applied
Physics	acoustics applied atomic molecular chemical cond matter fluids plasmas mathematical mechanics multidisciplinary physics nuclear particles fields astronomy astrophysics
Others	automation control systems nanoscience nanotechnology neuroimaging neurosciences nuclear sciences tech operations research management science optics polymer science robotics sport sciences statistics probability telecommunications thermodynamics transportation science technology zoology

Table 5-B: List of the 50 research categories retrieved from the WOS data set.

WOS data set simply by replacing PACS codes with numeric codes associated with the research categories listed in Table 5-B. The two datasets complement each other by providing both broad and detailed information about scientific production respectively at the macro level of natural sciences (WOS) and at the micro level of physics (APS).

5.1.1 Construction of the networks

Drawing on the two data sets, I constructed the co-authorship and the citation networks. In the co-authorship network, each node is an author, and two authors are linked if they published at least one article together. The network is described by the adjacency matrix $A = \{a_{ij}\}$, where entry a_{ij} is equal to 1 if author i and author j have co-authored at least one article, and is 0 otherwise. In the citation network, each article is a node, and a directed link is established from article i to article j if article j appears in the bibliography of article i. The citation network, combined with the information on the authors of each article, allows to associate citations directly to authors, and to compute the number of citations $N_i^{cit}(t)$ received by each author i over time.

5.2 Quantifying interdisciplinarity and success

Data about the subdivision in research categories and the number of citations obtained by articles are very suitable to investigate the impact of interdisciplinary research on author's performances. In this section I operationalise the concept of interdisciplinarity at the author's level and I introduce and distinguish between the two concepts of *background interdisciplinarity*, and the *social interdisciplinarity*. I illustrate a citation

rescaling procedure introduced in [115] useful to compare authors with different career lengths and to account for the various citation patterns that characterise distinct scientific domains. The various measures here defined are then used in Section 5.4 to shed light on the interplay between authors' interdisciplinarity and scientific performance. For each author i, I tracked over time the evolution of several measures: the background entropy $B_i(t)$ and the social entropy $S_i(t)$ at year t of author i's career; the number of published articles; the cumulative number of citations $N_i^{cit}(t)$ received up to year t; the normalised number of citations $\tilde{N}_i^{cit}(t)$.

5.2.1 Rescaling authors' careers and citations

Since the data sets contain authors who started their careers at different years, results may be affected by biases related to differences in career lengths. Indeed, authors who started their careers near the end of the data set (i.e., 2014) may have a smaller degree of interdisciplinarity and smaller number of citations than authors with longer careers simply as a result of, respectively, the smaller number of articles they have published, and the short time interval in which these articles could acquire citations. In order to address these problems, I first rescaled all careers to a common starting time t_0. In this way, a senior professor and a young researcher can be compared by considering them at the same time during their careers. I refer to the calendar time as $\tau \in [1945, 2014]$. The rescaled time t used in the analysis is then $t = \tau - \tau_{start}$, where τ_{start} is the calendar time of an author's first publication. Additionally, articles concerning with different areas of research may obtain different number of citations, not only because of their intrinsic impact and quality, but also as a result of the different citation practices and

traditions of various fields. Indeed, Radicchi et. al. have shown in [115] that the probability that an article obtain a certain number of citations has large variations between different disciplines, and it is widely recognised that comparing bare citation numbers is inappropriate. In order to compare adequately the performances of authors with different career lengths, working in different research fields, and who have started their career at different point in history I have adopted the rescaling procedure proposed in [115]. This method allows to account for variations in: (i) patterns and volume of citations across sub-fields and disciplines; (ii) attractiveness of research topics over time; and (iii) the starting year and duration of authors' careers. First, I computed the average number of citations $N_0^{cit}(\tau, c)$ received up to 2014 by all articles published in a given year τ and associated with a given PACS code or WOS research category c. For each article a, I then divided the total number of citations N_a^{cit} obtained by a up to 2014 by the average number of citations $N_0^{cit}(\tau, c)$ obtained up to 2014 by all articles associated with the same code or category c and published in the same year τ as article a:

$$\tilde{N}_a^{cit} = \frac{N_a^{cit}}{N_0^{cit}(\tau, c)}. \tag{5.1}$$

Following [116], for each author i at each year t of career, I obtained the normalised number of citations $\tilde{N}_i^{cit}(t)$ as the sum of the normalised number of citations $\tilde{N}_a^{cit}$ received by each article a that author i published in each year up to t. So constructed, this measure captures the success of an author as his/her relative performance in comparison with other similar authors.

5.2.2 Background entropy

I use the PACS codes and research categories (respectively, for the APS and WOS data sets) associated to the articles of an author to identify the author's research interests and expertise. In order to measure author i's interdisciplinarity [117], I construct the set $\boldsymbol{PC}_i(t)$ of *personal codes or categories*, defined as the collection of PACS codes extracted from all the articles published by author i up to year t. The set $\boldsymbol{PC}_i(t)$ thus reflects the disciplinary areas to which author i has contributed, and can be used as a proxy for i's (cumulative) background knowledge [117, 118]. I measure author i's *background interdisciplinarity* through the *background entropy* defined as the Shannon entropy of the set $\boldsymbol{PC}_i(t)$ of the author's personal PACS codes or research categories [119]:

$$B_i(t) = -\sum_{\alpha} p_i^{[\alpha]}(t) \log(p_i^{[\alpha]}(t)), \tag{5.2}$$

where the sum runs over all classes of codes in $\boldsymbol{PC}_i(t)$, $p_i^{[\alpha]}(t) = \frac{n_i^{[\alpha]}(t)}{\sum_\beta n_i^{[\beta]}(t)}$ is the probability of finding PACS code α in $\boldsymbol{PC}_i(t)$, and $n_i^{[\alpha]}(t)$ is the number of times a given PACS code α appears in $\boldsymbol{PC}_i(t)$. Similar entropy-based measures have been used for quantifying the heterogeneity of the citations made by an article [107, 108]. In general, authors with a more heterogeneous background are characterised by higher values of B, whilst smaller values are typically associated with authors whose research is focused on a small number of scientific sub-fields or categories.

5.2.3 Social entropy

By forging collaborations, scientists are exposed to various sources of knowledge, which may not be entirely coextensive with their own personal background, and on which they can rely to widen the scientific horizons of their research. To assess scientists' exposure to their collaborators' knowledge, I propose a measure that is meant to directly capture the social roots of interdisciplinarity. I first define the set $\boldsymbol{SC}_i(t)$ of *social codes or categories* of author i as the union of the sets of personal PACS codes associated with i's collaborators at the time of their last collaboration with i, from which I then removed the personal PACS codes of author i. So constructed, this measure takes into account only the PACS codes to which author i has been actually exposed during his career. Indeed if two authors i and j had published an article at time t but after t did not work together any longer, they would have not been reciprocally exposed to the PACS codes included in the articles each of them published separately after t. If instead the two authors had joined forces again and published a new article, for example after 5 years since t, then their sets of social PACS codes would have been updated and would include also the PACS codes associated with the articles they published independently during the 5-year period. I then measure the *social interdisciplinarity* of author i through the *social entropy* $S_i(t)$ defined as the Shannon entropy of the set $\boldsymbol{SC}_i(t)$:

$$S_i(t) = -\sum_{\alpha} q_i^{[\alpha]}(t) \log(q_i^{[\alpha]}(t)), \tag{5.3}$$

where the sum runs over all classes of PACS codes in $\boldsymbol{SC}_i(t)$, $q_i^{[\alpha]}(t) = \frac{m_i^{[\alpha]}(t)}{\sum_\beta m_i^{[\beta]}(t)}$ is the fraction of PACS code α in $\boldsymbol{SC}_i(t)$, and $m_i^{[\alpha]}(t)$ is the number of times PACS code α

appears in $\boldsymbol{SC}_i(t)$. A drawback of entropy-based measures is the degeneracy occurring at values of entropy equal to zero, which may correspond to different configurations of PACS codes or research categories. Indeed, a value of Shannon entropy equal to zero may correspond to an arbitrary number of identical codes, and therefore it does not distinguish between an author with many articles all associated with the same unique code and an author with just one article and one code. A similar degeneracy is obtained in the case of a perfect uniform distribution of codes which, however, never occurs in our data sets. For this reason, only values of entropy equal to zero were excluded from the analysis.

5.3 Empirical results

Figure 5.1 reports the normalised number of citations $\langle \tilde{N}_i^{cit}(t) \rangle$ averaged over authors characterised by a certain value of background entropy at four career stages, namely at $t = 5, 10, 15, 20$ years, in physics (APS data set, panels a-d) and in the natural sciences (WOS data set, panels e-h). On average, authors with intermediate values of background entropy (i.e., neither interdisciplinary nor specialised) are characterised by a relatively low value of scientific performance. Interestingly, both in physics (Figure 5.1(a-d)) and in the natural sciences (Figure 5.1(e-h)), scientific performance exhibits a U-shaped trend, with a minimum located in the range $B_i \in [0.2, 0.8]$, and two maxima in correspondence of the right and left extremes of the curves. Interestingly, a similar and relatively large number of citations can be obtained equally by highly specialised and highly interdisciplinary authors. However, the asymmetry in the U-shaped trend

indicates that, on average, scientists with the widest range of research interests ($B_i \gtrsim 1.2$) tend to outperform not only less interdisciplinary scientists, but also the most specialised ones ($B_i \simeq 0$). Overall, these results suggest that, at the micro level of physics as well as at the macro level of the natural sciences, both specialised and interdisciplinary scientists can be successful; yet extreme interdisciplinarity provides competitive advantage over extreme specialisation. While the U-shaped functional form is found already at the fifth year of a scientist's career, Figure 5.1 also suggests that the relationship between interdisciplinarity and success is subject to a temporal drift. At the macro level of the natural sciences (WOS data set), while the maximum normalised number of citations accrued by the most interdisciplinary author i at the fifth year of career is, on average, $\tilde{N}_i^{cit}(5) \simeq 160$, the largest value of $\tilde{N}_j^{cit}(20)$ for an author j at the 20-th year of career is, on average, $\tilde{N}_j^{cit}(20) \simeq 250$. Thus, an author i with, for instance, $B_i \simeq 1.0$ at the fifth year of i 's career will have, on average, 30 normalised citations, namely 0.40 times as many citations as those accrued by the most specialised author j (i.e., with $B_j \simeq 0$), and only 0.20 times as many citations as those of the best-performing author (i.e., with $B_j \simeq 1.40$). At the 20-th year of his or her career, the same author i with $B_i \simeq 1.0$ would still be able to accrue, on average, about 30 normalised citations. However, while i's comparative disadvantage over the most specialised author would remain unaltered, the disadvantage over the most interdisciplinary one would further deteriorate. Author i would therefore need to keep increasing background interdisciplinarity over time, lest by the 20-th year of i's career the total number of citations be, on average, only 0.12 times as large as the one of the best-performing author. The effects of background interdisciplinarity on success thus become more pronounced as scientists' careers progress. Moreover, in the long

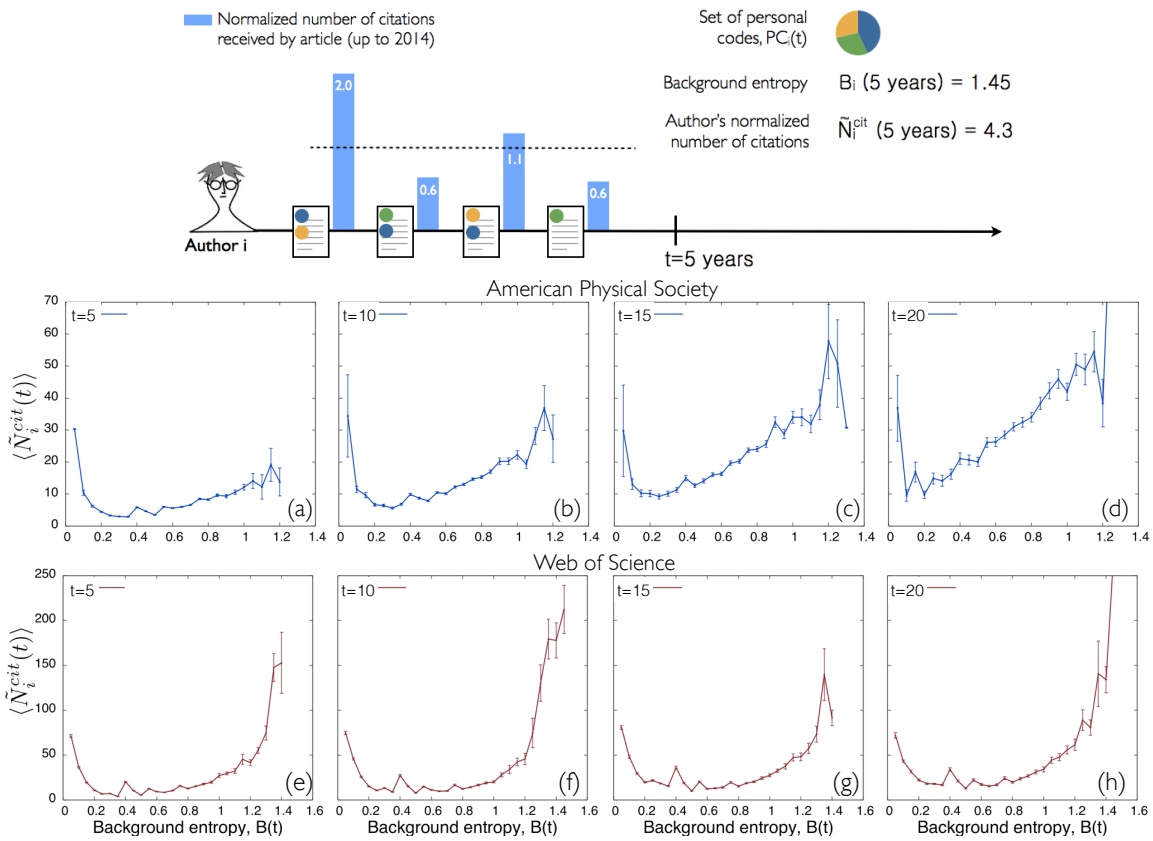

Figure 5.1: **Background interdisciplinarity and success at different career stages.** The normalised number of citations obtained by authors at different career stages as a function of their background entropy in physics (panels a-d) and the natural sciences (panels e-h). The vignette illustrates how performance and interdisciplinarity were measured. For each author i and a given year t of i's career, both performance and interdisciplinarity were measured at t on all articles published by i since the beginning of i's career up to t. The citations accrued by each article up to t were normalised through the method proposed in [116]. The U-shaped dependency of $\langle \tilde{N}_i^{cit}(t) \rangle$ on background entropy, and the presence of a minimum at intermediate values of $B(t)$ characterise both young authors and experienced ones, thus indicating that extreme interdisciplinarity as well extreme specialisation are already beneficial at the very beginning of a scientist's career. However, the competitive advantages of background interdisciplinarity become more pronounced as careers progress towards their final stages when the difference in performance between the most interdisciplinary and the most specialised authors reaches its peak. Error bars represent the standard error of the mean.

run, as careers approach their final stages, not only are highly interdisciplinary scientists more successful than specialised ones, but the difference in performance between the most interdisciplinary and the most specialised scientists reaches its peak.

Similarly, Figure 5.2 reports the normalised number of citations $\langle \tilde{N}_i^{cit}(t) \rangle$ averaged over authors characterised by a certain value of social entropy at four career stages, namely at $t = 5, 10, 15, 20$ years, in physics (APS data set, panels a-d) and in the natural sciences (WOS data set, panels e-h). Results indicate that authors can amplify success as their social interdisciplinarity increases and that the impact of social interdisciplinarity is almost linear across all stages of a scientist's career. An author with a more heterogeneous network (i.e., a higher value of social entropy S) will, on average, have a higher performance than an author with a more homogeneous network (and lower S). Thus, while specialisation can be a successful strategy (Figure 5.1), seeking collaborators with few and overlapping specialities will be a hindrance. Scientists can instead enhance their performance by selecting collaborators who are interdisciplinary or specialised in many different areas of science. Robustness checks based on null models and based on different measures of success are presented in the Appendix.

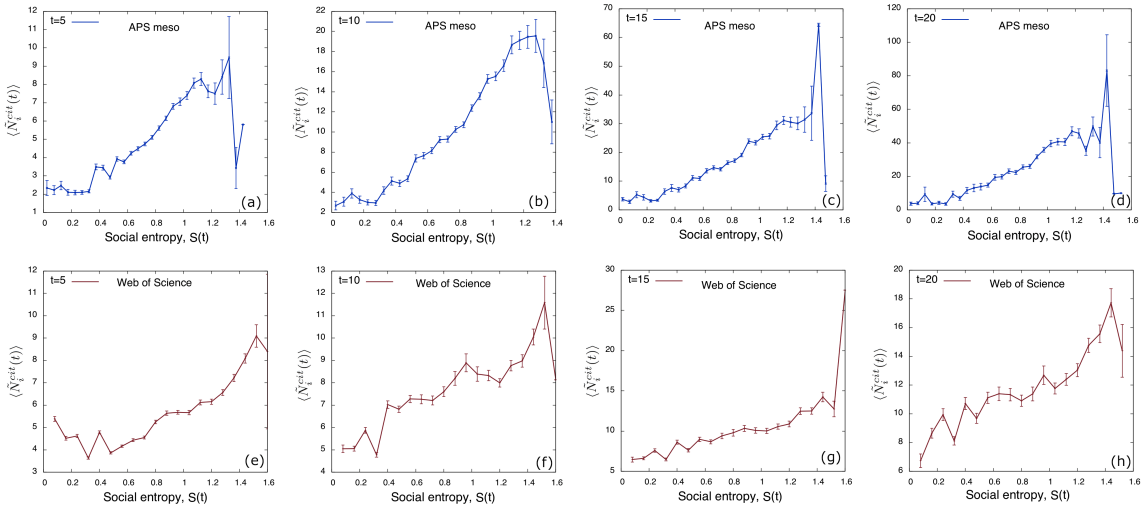

Figure 5.2: **Social interdisciplinarity and success at different career stages.** The normalised number of citations $\langle \tilde{N}_i^{cit}(t) \rangle$ obtained by authors at different career stages as a function of their social entropy in physics (panels a-d) and the natural sciences (panels e-h). For each author i and a given year t of i's career, both performance and social interdisciplinarity were measured. The linear and positive impact of the social entropy on the normalised number of citations characterises both young authors and experienced ones, thus indicating that a network of collaborators which provides more heterogeneous knowledge is beneficial since the very beginning of a scientist's career.

5.4 Path to interdisciplinarity

Given the advantages of interdisciplinarity, how do scientists widen their background over time, and which collaboration strategies are associated with success? I identify three collaboration strategies through which authors can enrich their set of personal codes: the solo, the absorptive and the synergistic strategy. First, I define the *solo strategy* as the acquisition of new knowledge through the publication of a single-authored article in the corresponding scientific area. With this strategy, scientists extend their background interdisciplinarity through "in-breadth" learning; yet they do not amplify their social exposure to new sources of knowledge. Second, the *absorptive strategy* is defined as the acquisition of new knowledge through the publication of a multi-authored article with at least one co-author who has already published in the corresponding area. Through this strategy, scientists absorb knowledge from their collaborators as soon as they are exposed to it, thus increasing their background interdisciplinarity (and possibly the heterogeneity of their collaboration networks). Finally, the *synergistic strategy* is defined as the acquisition of new knowledge by an author through the publication of an article with co-authors who have never published in the corresponding area. Through this strategy, collaboration promotes cross-fertilisation of various disciplinary areas, and ultimately intensifies all co-authors' background interdisciplinarity through the acquisition of new knowledge. An illustration of the three strategies is reported in Figure 5.3(a). In order to track exchanges and acquisition of new knowledge I have used a finer level of the PACS classification scheme, (i.e., the second hierarchical level). This choice is justified by the fact that: (i) individual articles usually concern with a very specific piece of knowledge and, (ii) during the joint production of an article authors can't and absorb an entire

discipline, but just the specific topic and problem addressed by the article as identified by PACS codes. I denote by P_i^{solo}, P_i^{abs} and P_i^{syn} the fraction of PACS codes acquired by author i through, respectively, the solo, absorptive, and synergistic strategies, during i's entire career. To understand how authors with different performance vary in their usage of the three strategies, in Figure 5.3(b) I show the average frequencies of solo, absorptive, and synergistic strategies adopted by authors in the APS data set whose articles accrued a total number of citations exceeding various thresholds. Remarkably, the overall frequency of the solo strategy is just about 4% at all levels of success, whilst the vast majority of the new PACS codes (about 96%) originate from collaboration. In particular, not only are authors across all levels of performance more likely to embrace a new sub-field through a synergistic strategy than an absorptive one, but also the difference in usage frequencies between the two strategies widens as authors are more successful ($N^{cit} \sim 10,000$). Exposure to collaborators' knowledge may broaden a scientist's background interdisciplinarity not only instantaneously through the absorptive strategy. When engaged in a joint endeavour, scientists can, in principle, gain access to the entire spectrum of knowledge offered by their collaborators [53]. A fraction of this knowledge can indeed be absorbed as soon as collaboration occurs, through the absorptive strategy; the remaining can be acquired at a subsequent stage, through a process here referred to as *postponed absorption*. Diluting acquisition of new knowledge over time can have various effects on scientific performance, depending on how much knowledge is acquired and on the time separating acquisition from exposure. To study the degree to which knowledge acquisition is affected by past collaborations, for each author I quantify the propensity, once exposed to a new social PACS code α, to acquire it at some subsequent

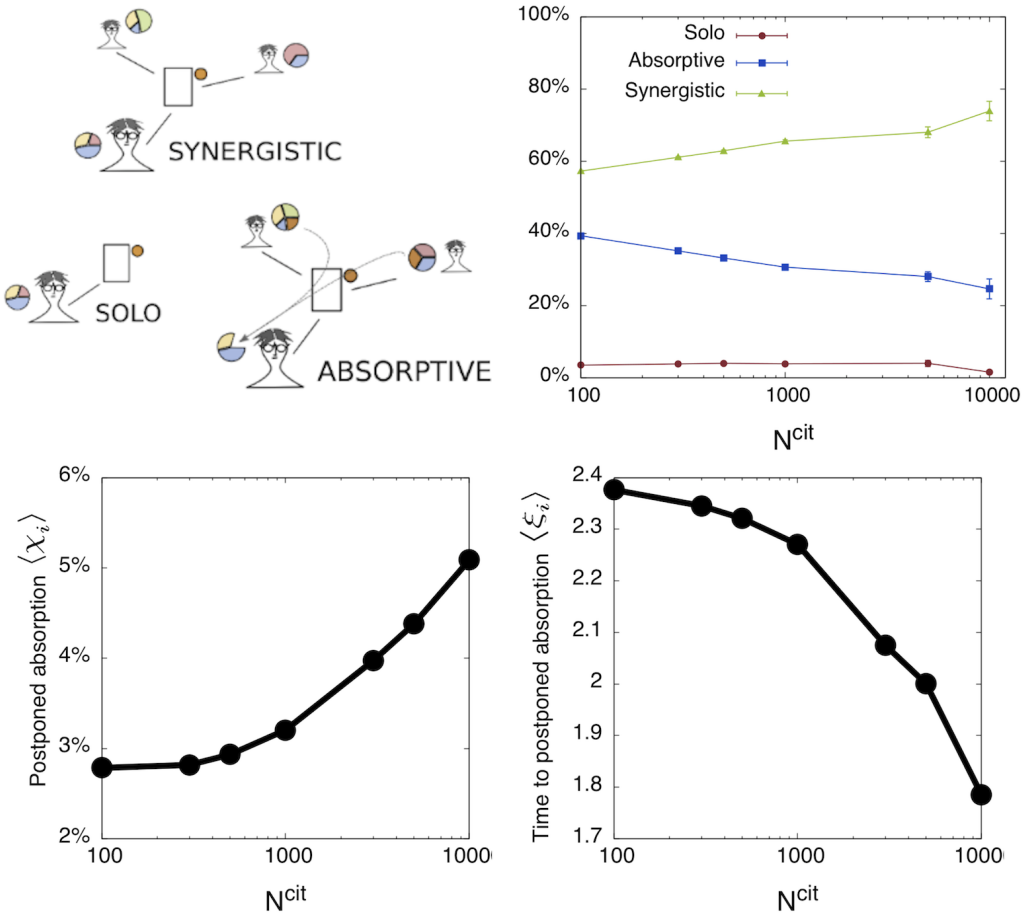

Figure 5.3: **Path to interdisciplinarity** (a) The three main strategies through which authors can expand their knowledge into a new field: (i) by publishing on their own in the new field (solo strategy); (ii) by collaborating with others that have already published in the field (absorptive strategy); and (iii) by collaborating with others that have never published in the field (synergistic strategy). (b) For authors in the APS data set whose articles obtained more than a given number of citations, I measured the average frequencies of solo, absorptive and synergistic strategies. Successful authors are more prone to synergistic strategies than less successful ones. Error bars represent the standard error of the mean. (c) The average fraction of social PACS codes eventually acquired by an author is positively correlated with the author's success. (d) The average time needed to acquire new PACS codes from collaborators is negatively correlated with an author's success. Successful authors are more likely not only to acquire knowledge from their collaboration network, but also to do so more quickly than less successful ones.

stage. Exposure to sub-field α occurs when author i, with no experience in α, for the first time collaborates with someone who has already published in α. Postponed absorption of α occurs when, for the first time after exposure, i appears as the solo author or co-author of an article a in α. Notice that the co-authors of a are assumed not to have experience in α (or else knowledge acquisition would be classified as instantaneous absorption). Of the social PACS codes to which author i was exposed, I measure the fraction χ_i that was eventually acquired by i. Lastly, for each author i I measure the mean interval of time ξ_i separating postponed absorption from exposure. Figure 5.3(c) shows the average fraction $\langle \chi_i \rangle$ over all authors, and suggests that successful authors in the APS data set are more likely to use up their collaboration networks to acquire new knowledge than less successful ones. Moreover, Figure 5.3(d) shows the average interval of time $\langle \xi_i \rangle$ over all authors, and suggests that the time separating exposure to new knowledge from acquisition tends to become shorter as authors' performance increases. Not only do successful scientists choose their collaborators carefully so as to secure exposure to new areas, but they also prefer not to wait too long before they publish in those areas either on their own (solo strategy) or with other collaborators (synergistic strategy).

5.5 Discussion

Empirical findings suggest that highly heterogeneous personal knowledge significantly impact on authors' performances, and on their ability to produce high-impact research. However, scientists bear opportunity costs as they begin to diversify their background, at least until they become highly interdisciplinary. I also found that scientists benefit from

heterogeneous collaboration networks, corroborating the hypothesis in Section 2.3. The findings are also in agreement with previous work which has suggested that an individuals performance is attributable not only to competence, but also to the network in which the individual is embedded [6, 62]. Cognitively diverse networks that offer opportunities of knowledge recombination have been found to sustain innovation and knowledge creation [54, 55, 63]. In a similar vein, our conception of interdisciplinarity extends beyond the boundaries of the scientists background to also include their collaboration networks. I suggested that scientists can integrate and extend in-breadth learning by widening the breadth of their social network. Scientists with groups of collaborators spanning many different areas are more successful than those with collaborators focused on one or few overlapping areas. These results are in agreement with the idea of scientific innovation as socially-aided heuristic search process presented in Section 2.3. In particular, the empirical finding confirms the importance of the interaction and exposure to others' knowledge and perspectives. Finally I have also identified the most effective strategies used by scientist to acquire access to and absorb variegate knowledge while avoiding the cost of cognitive overload.

Chapter 6

NetworkL: a python package for the longitudinal analysis of complex networks

Over the last 20 years a number of real-world complex systems, such as communication, transportation, biological and technological systems, have been studied from a network perspective, i.e., by regarding the system as a graph in which the nodes are the system components whereas the edges represent the interaction between them. Real-world processes, such as spreading of diseases [120–122], traffic congestions [123, 124], human mobility and behaviour [125, 126], information routing [127, 128], have been extensively investigated by using network metrics already available *on the market* (e.g., centrality measures) or by proposing new ones [40, 113, 129]. More recently, a great effort has been devoted to the study of time-varying graphs [130], i.e., graphs in which nodes and edges

are created and deleted over time, and several new network metrics have been proposed in order to investigate and characterise appropriately the temporal nature of modern data sets [131, 132]. The analysis in chapter 4 and 5 of this thesis, and in general the Science of Success, make extensive use of the time dimension available in today's data sets. The importance of network analysis techniques in empirical investigation has stimulated the birth of a number of open-source software libraries for the manipulation, analysis and visualisation of networks. Among others, the most popular libraries for graph computation (widely adopted across many fields and outside the academia) are igraph [133], NetworkX [134], SNAP [135], and graph-tools [136]. To cope with the increasing size of electronic data sets some libraries and algorithms have been specifically designed to take advantage of modern multi-core processors or distributed hardware infrastructure [136]. For instance, the computational frameworks Giraph, GraphX, and Pragel, respectively by Facebook, Apache, and Google, have pushed graph analysis to the scale of billion of nodes and edges. However, in the era dominated by massive data sets and apparently unlimited computational power it is often mistakenly assumed that Big Graph requires Big Data approaches or extreme computational power. While the analysis of large amount of texts, images, videos requires, at least, massive and expensive storage solutions, the graph of one of the largest social network (Facebook, 1.55 billion active users in Q3-2015) could easily fit into a single commodity hard-drive when represented as an edgelist. Additionally, in a recent work [137], Kyrola et. al. have shown that extreme computational infrastructure may be unnecessary and overused in the domain of large graphs. In particular, in [137] it has been shown that careful improvements of existing approaches make it possible to use commodity hardware to perform computations

previously performed only on large-scale distributed hardware. If on the one hand such advances in graph algorithms make it possible to study graphs made of millions nodes on modern laptops, on the other hand new computational challenges are just around the corner. In recent years digital data are not just increasing in volume (e.g., number of nodes and edges in a graph) but are also extremely dynamic. The temporal dimension is ubiquitous in today's data sets from road traffic and public transport data, to the interaction on social and gaming platforms [138]. Additionally, temporal data are often available at a very fine-grained temporal resolution. As an example the Twitter graph (120 million users) changes in time at exceptional speed with thousands of follower relations created each second. By the time a simple PageRank [139] analysis is completed the Twitter graph has changed dramatically and the result of the computation would be out-of-date. If the analysis involves the computation of graph distances, as in the case of betweenness or the closeness centrality (used in Chapter 4), a real-time update of the results is beyond the capabilities of any existing library.

These computational limitations substantially prevent the exploration and exploitation of longitudinal data sets at the finest temporal resolution and have become today a concrete obstacle in the daily activity of both network researchers and businesses operating in the domain of data science. In particular, all works that involve the recomputations of certain network metrics over time need to trade-off computational speed against time resolution. In order to save computation time, data available at daily resolution may get aggregated to monthly or yearly snapshots, losing information hidden at smaller timescales. In this Chapter I introduce *NetworkL*, a Python package precisely devoted to help researchers and data scientists to unlock information hidden at smaller

timescale, using commodity hardware even for large graphs.

NetworkL manipulates dynamic graphs at the smallest time scale possible, i.e., the single-edge addition/removal, and dramatically facilitates the computation over time of distance-based network metrics by implementing and building-on incremental graph algorithms for shortest paths. The package primarily deals with the problem of continuously updating the values of all shortest path lengths upon the arrival of new edges over time. Such problem is encountered in the dynamic computation of several network measures such as global efficiency [140], closeness [141], betweenness [40, 142], and information centrality [129]. In particular it finds direct application in the re-computation of the closeness centrality in the WorldWide Startup network presented in Chapter 4.

Incremental graph algorithms date back as far as 1991 [143] when highly dynamic data sets such as the Twitter or the Facebook graph did not yet exist. Surprisingly, little or no slipover into the domain of network science has happened so far, as proven by the absence of any dynamic computation in the most popular software libraries [133–136]. A reason that may have limited the wide adoption of dynamic computation on graph is the fact that some incremental graph algorithms can be memory voracious. The greater speed at which computation is performed and results updated comes at the price of larger amount of memory required. As an example a static algorithm for computing the closeness centrality takes only the graph (e.g., adjacency list) as input and gives a N-dimensional vector as output. Notice that each entry of the geodesic matrix D is directly computed during the execution, but only the aggregation of its rows or columns is required to be stored in memory. Contrarily, an incremental approach to the same problem would take as input not only the original graph and the changes (i.e., the edges to

be added or removed), but also additional information about the solution at the previous time step, i.e., the geodesic matrix D. During the execution of the incremental algorithm each individual entry of the matrix D can be subjected to changes and consequently the full matrix needs to be available in memory adding a $O(N^2)$ space complexity. Such space complexity affects distance-based metric such as closeness of betweenness centrality and is the main reason why some recent implementations [144–147] heavy relied on large distributed hardware.

A novel contribution of my work is the introduction of a memory efficient strategy, based on what I call *Sparse Biconnected Geodesic Matrix* (SBGM), tailored for incremental computation, and able to save from 50% to 80% of the memory required by previous approaches in the case of real-word graphs. NetworkL and the SBGM approach make it possible to take advantage from incremental graph algorithm without the need to deal with the complexity of distributed hardware. The possibility to perform incremental computation on commodity hardware with limited memory is likely to widened the scope of empirical investigations to highly dynamic temporal networks. The success and improvement of tools such as NetworkL are definitely determined by the contribution of the developers and the network researches communities. My aim is to provide with this contribution a good starting point to facilitate empirical investigation and longitudinal analysis of graphs. Contributions to the definition of the future roadmap as well as code contributions on the GitHub project page `http://github.com/networkl/` are welcome. This chapter is organised as follows. Section 6.1 provides a quick overview on incremental graph problems, on the relevant literature, and on the most recent advancements in large scale graph computation. Section 6.2 introduces the SBGM and presents performance

testing on real-world network data sets. Section 6.3 describe the methods, functions and structure currently implemented in NetworkL and how to use them. Section 6.5 is devoted to conclusions, limitations, and the future roadmap.

6.1 Incremental graph problems

Generally speaking an algorithm can be described as a function f which, given an input g produces the output (the solution) $f(g)$. As an example, the computation of PageRank takes as input a graph g with N vertices and the function f mainly involves the inversion of an appropriate matrix obtained from the adjacency matrix provided as input. The output $f(g)$ is a vector of size N whose entries are the PageRank scores. In this example the algorithm can regarded as "static" because, once the graph is subjected to a change in the structure, the computation has to be performed entirely from scratch by applying f to the new modified input. However, if the modified input g' is not very different from the original graph g it may happen that the solution $f(g')$ is close to $f(g)$ and it can be computed without the need to apply again the function f. The simplest example of this fact is the re-computation of topological distances in a unweighted graph after the addition of a new node connected only by one new edge which is illustrated in Fig. 6.1. It is straightforward to notice that the new node i will inherit the distances to all the other nodes from its neighbour j and there is no need to apply the Dijkstra algorithm [148] to obtain the update solution. In particular, if d^g_{jk} is the topological distance between node j and a node k in the graph g, then the distance between i and k in the updated graph g' can be easily written as $d^{g'}_{ik} = d^g_{jk}+1$. Notice that, in order to obtain the updated solution

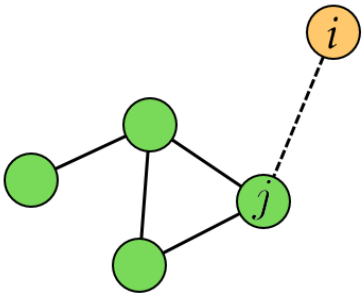

Figure 6.1: Example graph to illustrate the key concept of incremental graph problems.

$f(g')$, we have not used the steps encoded in the function f which describes the Dijkstra algorithm. An *incremental algorithm* can be described as a function ϕ which takes as input only the original input g, the output $f(g)$ and possibly auxiliary information about the changes $\Delta g = g' - g$, and provides the output $\phi(g, f(g), \Delta g) = f(g')$. The crucial point here is that f and ϕ can be in general very different and, more importantly, the computational complexity of ϕ may be smaller than f.

The characterisation of incremental problems, both in the domain of graph computation and not, has intrigued scholars for decades [145, 146, 149–154] and studies goes well beyond the illustrative example reported in Fig. 6.1. One important contribution in this field is the article by Ramalingam and Reps [143] in which a novel approach to assess and compare the computation complexity of incremental algorithms is proposed. In the article authors highlight the fact that expressing the cost of the computation as a function of the size of the input (e.g., the number of nodes in a graph) is not very informative in the case of incremental algorithms and one should instead characterise the complexity with respect to the amount of changes in the input Δg and in the output $\Delta f = f(g') - f(g)$. From this angle, the authors study the problem of updating shortest path distances on graphs and propose an incremental algorithm which is still largely

adopted in recent software implementation [155].

More recently, several works have proposed alternative strategies for the incremental computation of shortest distances [154, 156] as well as the incremental computation of network metrics such as betweenness centrality [155], closeness centrality [145], and PageRank scores [157, 158] in dynamic graphs. Often these recent software implementation rely on large-distributed computational infrastructure. When dealing with tens of Terabytes of data, distributed computational frameworks such as Hadoop [1] are extremely useful to speed up data-mining tasks (counting, aggregation, clustering, machine learning) which fit the parallelization and map-reduce paradigm. However, in the area of graphs several drawbacks are hidden in distributed computing. Among these is the need of passing information across the vertices of the graph and consequently across nodes in the computational cluster. The problem of minimising the message passing across the cluster relates to the problem of partitioning the graph and find efficient graph cuts, which is a hard problem [159]. Moreover, since the computation complexity of many graph algorithms is not linear and the growth rate of modern data is exponential, improving computational speed by scaling the hardware seems not a sustainable solution both economically and environmentally. My contribution embraces the recent trend towards the improvement of the computational efficiency on single machine [137] rather than increasing execution time of existing algorithms through the hardware scaling.

[1] `http://hadoop.apache.org/`

6.2 Sparse Biconnected Geodesic Matrix

As mentioned in the previous section the main drawback of the incremental computation of distance-based network metrics is the need to store in memory the full geodesic matrix D. Maintaining in memory the topological distances between all pairs of nodes is inefficient, unfeasible on commodity hardware in the case of medium-sized networks ($\sim 10^6$ nodes), and in general is not a scalable solution as it requires $O(N^2)$ space. Nevertheless a large body of literature overlooks this problems and overcomes the technical limitations by relying on large distributed memory resources. I present here the three strategies used in NetworkL to reduce drastically the amount of memory needed during incremental computation of shortest path distances. The common underlying idea is to drop unnecessary or redundant entries stored in the full geodesic matrix D while maintaining efficient and fast access to the value of the distance between all nodes pairs. I use the example network in Fig. 6.2 to briefly illustrate the strategies. Panel (a) shows the full geodesic matrix D while panels (b-d) show the different matrices obtained by removing redundant entries. The empty white areas in the three matrices in panels (b-d) provide also a visual representation of the amount of memory saved by each of the three strategies.

Sparse Geodesic Matrix. In the first strategy the matrix D is replaced by a sparse matrix, the Sparse Geodesic Matrix (SGM), in which the entries equal to the most frequent value of distance are dropped from memory. The most frequent distance in the example network is $d = 1$ (see Fig. 6.2(a)) and the corresponding entries are dropped in the SGM as indicate by the empty white areas in Fig. 6.2(b). The value $d = 1$ is

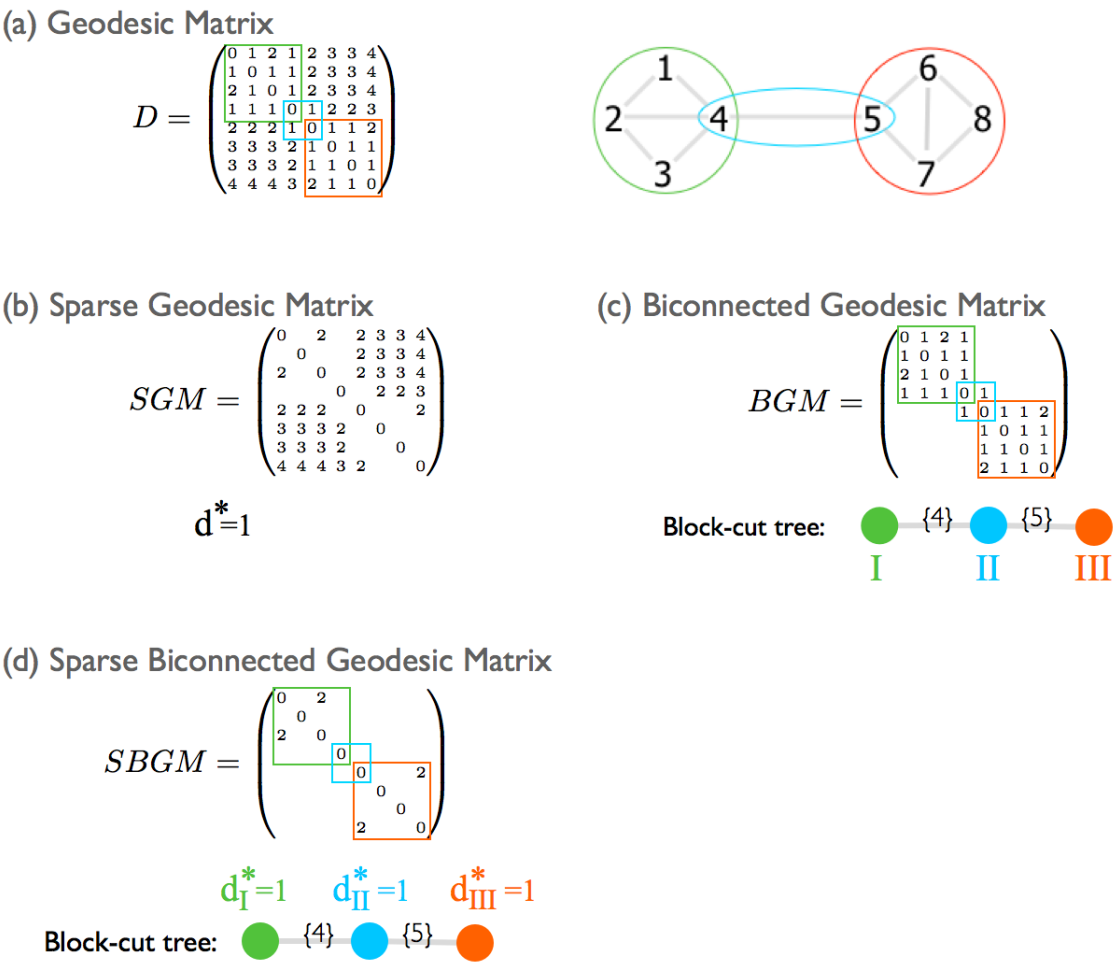

Figure 6.2: (panel a) Example network and its corresponding full geodesic matrix D. The three bi-connected components are highlighted by coloured circles in the graph and by rectangles respectively with the same colour in the matrix D. (panel b) The SGM saves the memory required to store the most frequent distance $d^* = 1$ found in the full matrix D. In the SGM implementation this value is stored only once in a separate variable which is returned when a missing entry (e.g., d_{12}) is required by the code execution. (panel c) The BGM stores only distances between nodes belonging to the same bi-component. Additionally a block-cut tree is stored in memory to assist the retrieval of the value of the distances between nodes in different components (i.e., the missing entries). (panel d) The two approaches combine in the BSGM in which several independent variables d^* are stored, one for each bi-component separately.

stored only once into a single scalar variable d^* instead of saving it multiple times in the matrix D. The SGM strategy achieves a significant memory reduction on real-world networks as they are often characterised by small diameters and picked distribution of nodes distances as show in Fig. 6.3. As an example the pick at $d_{ij} = 3$ in the *Wiki-vote* network [160] reveals that more than 50% of the entries of the geodesic matrix have the same value. The SGM associated with the *wiki-Vote* network thus occupies half of the memory required by the full geodesic matrix as shown in the comparison chart in Fig. 6.4. The current version of NetworkL (v.0.1) implements the SGM as a python object containing the following instance variables: (i) an integer *d-star* which is set equal to the value d^*, and (ii) a 2-levels nested python dictionary in which the keys correspond to the nodes in the network. At the fist level the dictionary contains as many keys as the number of nodes in the network, while at the second level the nested dictionaries include only those keys whose values differ from d^*. If a certain pair of dictionary-keys (i, j) does not exist then the value stored in the variable *d-star* is returned. Notice that all the diagonal elements d_{ij} need to be store explicitly even if they are all equal to zero by definition. In general the SGM can be implemented by using any existing sparse matrix representation and by considering the matrix $D' = D - d*$. The correct node distances can be subsequently retrieved as $d_{ij} = d'_{ij} + d*$.

Biconnected Geodesic Matrix. The second strategy is based on the decomposition of the graph into bi-connected components (bi-components). Such decomposition has been recently used [161] to speed up the (static) computation of betweenness centrality. Notably, the partitioning of a graph into bi-connected components can be obtained in linear time with existing algorithms [162, 163] and does not affect significantly the com-

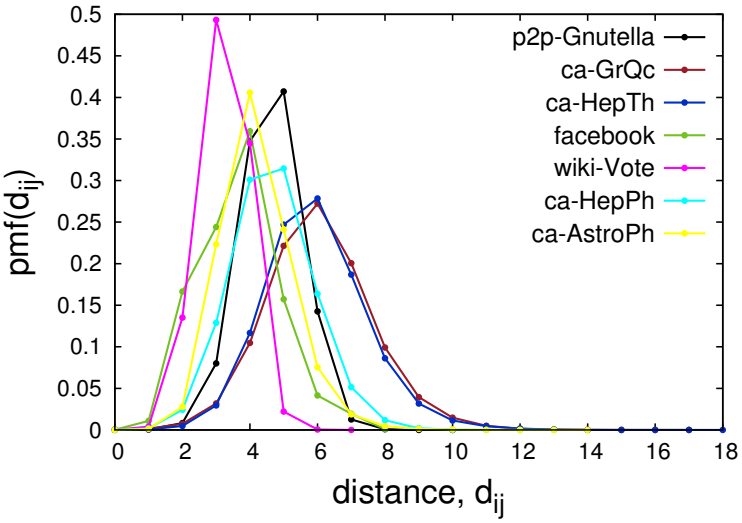

Figure 6.3: Distribution of node distances in several real-world networks. Notice that the number of entries in the geodesic matrix which are equal to the most frequent distance goes from a minimum of 25% in the *ca-HepTh* network to a maximum of 50% in the *wiki-Vote* network. This percentage corresponds to the amount of memory saved by the SGM strategy as shown in Fig. 6.4

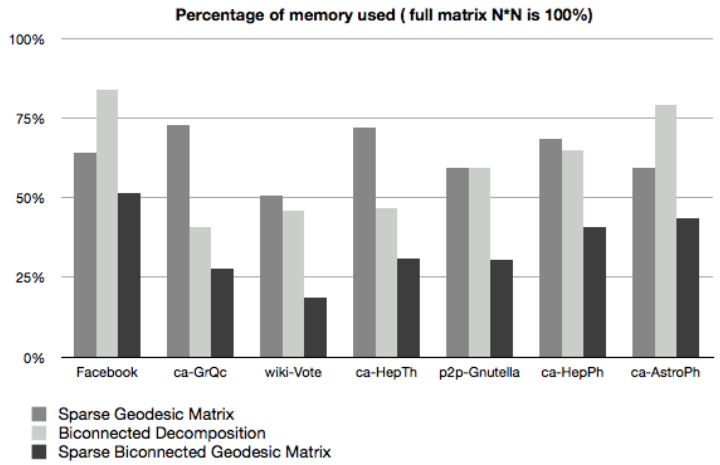

Figure 6.4: Comparison of memory reduction achieved by the three strategies.

putation performance. A biconnected component is a maximal biconnected subgraph, i.e., a subgraph which can not be broken into disconnected components by removing one vertex only. The example network in Fig. 6.2(a) contains 3 bi-components identified respectively by three coloured circles. Nodes 5,6,7,8 form a bi-component because it is

not possible to make any two of them unreachable by removing only one of the nodes. On the contrary, the removal of node 4 (or 5) splits the graph into two disconnected components. Any connected graph decomposes into a tree of biconnected components called the block-cut tree of the graph (see Fig. 6.2(c)). The blocks are attached to each other at shared vertices called cut vertices or articulation points (i.e., the nodes 4 and 5). The removal of an articulation point splits the original graph into multiple disconnected components. The block-cut tree shown in Fig. 6.2(c) makes it clear that all shortest paths between the nodes in adjacent bi-components must cross the associated articulation point. For instance in order to go from node 1 to node 5 it is necessary to cross the articulation point $\{4\}$ which connects the two blocks. Additionally, the path between node 1 and node 8 can be built as a combination of the paths from node 1 to the articulation point 4, from node 4 to node 5, and from the articulation point 5 to node 8. For this reason it is unnecessary to store all the entries d_{ij} of the geodesic matrix whose nodes i and j belongs to different components. In this way the geodesic matrix can be reduced into a "sort of" block diagonal matrix, the Biconnected Geodesic Matrix (BGM) in which each block corresponds to a biconnected component and two blocks overlap in correspondence of the articulation points. Notice also that the BGM approach can be used on weighted graph without modifications of the procedures.

Figure 6.2(c) shows the BGM obtained for the example network. As before the empty white areas indicate the regions of the matrix which are not actually stored in memory. One advantage of the BGM is that when a new edge is added between nodes belonging to the same bi-component (e.g., between nodes 5 and 8) only the shortest paths within the same bi-component can be affected, while the entries in other blocks remain unchanged.

More importantly, while the new edge $(5, 8)$ effectively reduces the distance from node 8 to nodes $\{1, 2, 3, 4, 5\}$ and requires to decrease 10 entries in the full matrix D, the same edge addition induces the decrease of only 2 entries in the BGM, namely d_{58}^{BGM} and d_{85}^{BGM}, because all other distances will continue to be inherited from the articulation points. The absence of entries in the off-diagonal regions of the BGM provides a huge saving in terms of number of reading/writing operations from/to memory when the size of the graph and the number of bi-components increase. These advantages do not come without some limitations. First, BGM does not provide direct memory access to all entries as in the case of the SGM. The inter-blocks distances (e.g., d_{18}) require the computation of one single-source/single-target shortest path on the block-cut tree which, however, can be performed efficiently because of the absence of loops and which has a computational cost that can still compete with the cost of messages passing and graph partitioning across distributed computational clusters. Second, the BGM strategy makes necessary to store additionally the block-cut tree, which in the worst case requires $O(L)$ memory, and a one-to-one mapping between nodes an bi-components, which requires $O(N)$ memory. Lastly, the addition and removal of edges during the dynamic computation may alter the structure of the bi-connected decomposition and consequently the block-cut tree. An edge added between nodes in two adjacent bi-components creates a new bi-components which contains all nodes in the two original bi-components. As a result, the respective bi-component blocks in the BGM matrix need to be merged into one single block and some off-diagonal entries have to be added. If the new edge connects bi-components which are far apart in the block-cut tree then the new edge is introducing a loop in the tree and all blocks included in this loop must be merged into a single block. In the case

of an edge removal (not braking connectivity of the entire graph) the only effect can be a split of a bi-component into two distinct bi-components. This process has little impact on the block-cut tree structure, and allows to free additional memory, namely all the entries d_{ij} such that i and j are in the newly created bi-components. However, to the best of my knowledge, there are no algorithms to check the bi-connectivity of a subgraph in an incremental fashion after an edge removal, and any potential split has to be identified by the standard decomposition algorithm applied from scratch to the modified subgraph.

Figure 6.4 reports the huge saving that can be achieved by BGM on real-world networks. BGM outperform SGM in several data sets going from a minimum of about 20% memory saving (Facebook network) to a maximum of about 60% memory saving (ca-GrQc network). In the current version of NetworkL (v.0.1) the BGM is implemented as a python class containing the three following instance variables: (i) the block-cut tree implemented as a NetworkX object, (ii) a dictionary mapping the components IDs to which a node belongs, (iii) a 2-level nested dictionary representing the BGM and containing as many first-level keys as the number of nodes. The value of the latter dictionary are special python objects implemented by the *BiconnectedDict* class which inherit from the standard *Dict* class. When a second-level key is missing (e.g., the entry d_{18}) the *BiconnectedDict* object call a method in the BGM class which performs the shortest path computation on the block-cut tree and return the value of distance as the sum of intermediate values at the articulation points (e.g., $d_{18} = d_{14} + d_{45} + d_{58}$). This computation is performed transparently and the user can retrieve all distance values with the unique usual expression for python matrices: d[1][8].

Sparse Biconnected Geodesic Matrix. The two strategies can be combined together to achieve additional memory saving. The Sparse Biconnected Geodesic Matrix (SBGM) is a block matrix similar to the BGM in which each block is independently regarded as a sparse geodesic matrix. The network is first decomposed into bi-connected components and then we identify the most frequent distance value in each bi-component. As occurs with the SGM, the SBGM does not store in memory these frequent distances which are stored only once in as many variables $d^*_{\mathrm{I}}, d^*_{\mathrm{II}}, d^*_{\mathrm{III}}, ...$ as the number of bi-components. In the example reported in Fig. 6.2(d) the most frequent values in each component happen to be the same ($d^*_{\mathrm{I}} = d^*_{\mathrm{II}} = d^*_{\mathrm{III}} = 1$) for all blocks but in general the value of d^* may differ across bi-components. The visual representation of the SBGM in Fig. 6.2(d) shows that only 8 diagonal elements plus 4 inter-blocks entries need to be stored. Additionally, three d^* variables plus eight integers representing the edge-list of the block-cut tree need to be saved. The grand total sum up to 23 integers which is around 1/3 of the 64 integers stored in the full matrix D. Implementation details of the SGM and BGM are transferred directly to the SBGM which inherits strengths and limitations of both approaches while providing significantly better memory performances. SBGM achieves an astonishing 75% reduction of memory for the *Wiki-vote* network and does not go below 50% reduction in all other data sets studied (see Fig. 6.4).

6.3 Using NetworkL

NetworkL source code is released under GNU V3 license. It can be downloaded for free from the github project page `http://networkl.github.io/`, and the package can be

installed through the popular package manager *pip* by typing *pip install networkl*. The package has been primarily developed with the aim to perform incremental computation and rapidly test the ideas and potential of the sparse geodesic and the bi-connected geodesic matrices (SGM,BGM). For this reason extreme optimisation and performance were not the foremost concern. Pure python after all is not the primary choice for fast computation and the future roadmap already includes a C++ implementation. Nevertheless, the code has reached a level of maturity which makes it useful and preferable to static computations in many scenarios. In this section I briefly illustrate the code structure, the classes, functions and methods implemented and I provide example of usage.

Code structure is modular and includes 3 classes which implement the three matrices SGM, BGM, and SBGM, and the function *update-distance-matrix()* which implements the Ramalingam Reps algorithm for shortest path updating [143]. In the current released version (v0.1) only SGM is implemented, while BGM and SBGM are still in a local development branch. The function *update-distance-matrix()* takes 4 objects as input: the graph G (a NetworkX object), a geodesic matrix (alternatively full matrix D, SGM, BGM, or SBGM), an edge (i, j) and a string equal to "add" or "remove" with straightforward meaning. The graph G is modified in place meaning that during the execution of *update-distance-matrix()* also the graph object will be modified by adding (or removing) the edge (i, j). The updated geodesic matrix can be used subsequently to compute some distance based network metrics such as closeness centrality, graph efficiency [140], information centrality [129]. Future roadmap includes the possibility to compute incrementally these metrics by using a specific function, e.g., *update-closenness-centrality()*

which takes the old metrics scores as additional input.

The SGM is implemented in the *SparseGeoMatrix* python class which includes the d^* value and the sparse geodesic matrix as instance variables as well as an *optimize-dstar()* method which can be invoked opportunistically to identify the most frequent value of distance and modify the matrix accordingly. In the current version of NetworkL the *SparseGeoMatrix* class still lacks an automated and efficient strategy to invoke the optimisation of d^* which is left to future development. At the user level NetworkL provides also two more functions useful to take advantage of the SGM strategy. The function *geodesic-to-sparse-geodesic(D,G)* converts the full geodesic matrix D into a SGM (only if the graph G is connected). The function *sparse-distance-matrix(G)* takes a graph H as input and produces a SGM as output. This computation is still performed by constructing first a full geodesic matrix D and invoking then the *geodesic-to-sparse-geodesic(D,G)* function.

The BGM is implemented in the *BiconnectedGeoMatrix* python class (development branch) which includes as instance variables: the block-cut tree (a NetworkX Graph object), the mapping between nodes and bi-components, a nested dictionary representing the BGM, a pointer to the graph G. When a new BiconnectedGeoMatrix object is created the bi-connected decomposition algorithm is performed and the block-cut tree, the matrix and the mapping are initialised with the computed values. All-pairs shortest paths computations are then performed for each bi-component separately. Each value of the matrix instance variable contains a *BiconnectedDict* class which behaves as a standard python dict class with the only difference that if a certainentry is missing (e.g., the entry d_{18} in Fig. 6.2) then the *get-inter-block-distance()* method from the *Biconnected-*

GeoMatrix class is called to reconstruct the required value as the sum of intermediate distance values at the articulation points (e.g., $d_{18} = d_{14} + d_{45} + d_{58}$ in Fig. 6.2). Merging and splitting of components are still in development. The SBGM class will inherit from the *BiconnectedGeoMatrix* python class. The only difference is that each block will be treated as a sparse geodesic matrix. Implementation of the SBGM class is still in development.

6.4 Performance and testing

I have conducted a number of benchmarking tests to assess the computational speed of the *update-distance-matrix()* function. The aim of these tests is primarily to provide NetworkL users with a general idea of the typical computational time required for each edge addition/removal on real data sets. The aim is not yet to provide a faster tool for incremental computation (for which a C implementation would be more appropriate) but to test the performances of SGM and BGM during incremental computation and comparing them with the baseline of static computation already provided by NetworkX (v1.10). The improvement obtained on the python baseline suggests that also faster languages (C,C++) can benefit from the SGM and BGM approaches. The tests have been conducted on the seven network data sets reported in Fig. 6.4.

In the first test I start from a minimum spanning tree of the network and I add all remaining edges one by one until I obtain the original graph. After each edge addition

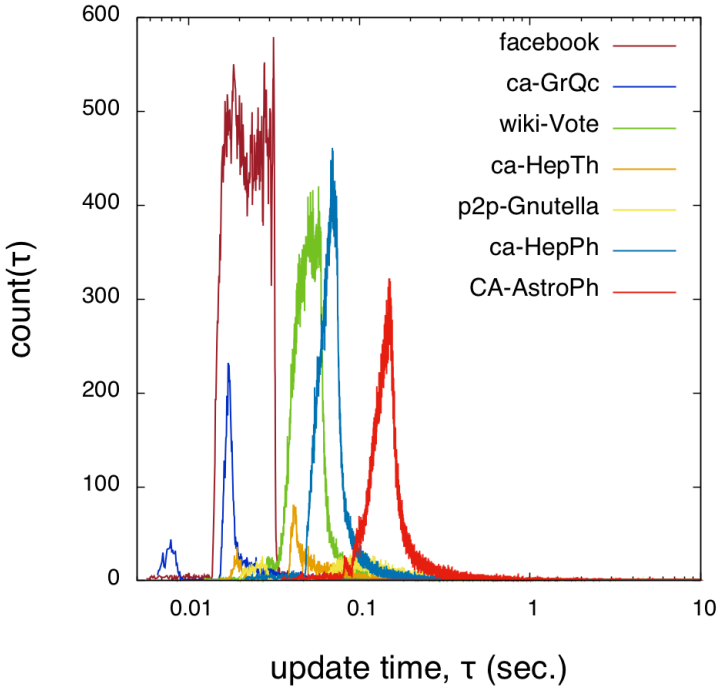

Figure 6.5: Frequency plot of the time τ needed to update the SGM after each individual edge addition for several network data sets. Remarkably the time needed to update the entries of the SGM is on average below 0.1 seconds for all data sets.

I invoke the *update-distance-matrix()* function giving as input the SGM, and I take record of its running time. In Fig. 6.5 I plot the distribution of updates times for the various networks. The figure shows that the distributions are usually picked around one or two values and that the average update time is below 0.1 sec. for all data sets studied. Table 6-A reports a comparison between the time needed to compute all-pairs shortest path *from scratch* with the Stanford SNAP library [135] and the maximum update time per edge addition obtained with the NetworkL library during the execution of the benchmarking tests. For almost all data sets the computation time *from scratch* is comparable with the maximum time needed by NetworkL to update all distances in the SGM. Moreover, the most frequent time needed to update the entire SGM is, on average, below 0.1 seconds (see Fig. 6.5) which is significantly smaller than the time

Data set	NetworkL max update time (sec.)	SNAP library [135] all-pairs shortest path time (sec.)
Facebook	37	47
ca-GrQc	21	14
wiki-Vote	613	117
ca-HepTh	63	64
p2p-Gnutella	174	125
ca-HepPh	311	240
ca-AstroPh	1245	713

Table 6-A: Comparison between the maximum update time per edge addition (NetworkL library) and the time needed to compute all-pairs shortest path *from scratch* with the SNAP library. For almost all data sets the computation time *from scratch* is comparable with the maximum time needed by NetworkL to update all distances in the SGM. Moreover, the average and most frequent time needed to update the entire SGM is below 0.1 seconds (see Fig. 6.5) which is significantly smaller than the static computation.

required by the *from scratch* computation.

I also tested the correctness of our implementation of the Ramalingam and Reps update algorithm. At the end of the each addition/removal of an edge I compare the entries of the updated SGM with the entries of the full geodesic matrix computed with the NetworkX (v1.10) function *shortest-path-length()*. About a total of 7×10^8 entries have been compared across the various data sets and no mismatch was found.

6.5 Conclusion and future development

In this Chapter I have introduced NetworkL, an python package for the dynamic computation of graph metrics on time-varying networks. NetworkL implements incremental algorithms for shortest path recomputation originally proposed by Ramalingam and Reps [143]. After the addition or removal of an edge in the graph the algorithm perform the

minimum set of operation to update the shortest path length without the needs to recompute those paths not affected by the edge change. The size of today's network data sets, made of tens of thousands of nodes and edges, poses significant challenges to maintain updated in memory all entries of the geodesic matrix required by the incremental computation. To overcome this limitation I propose three novel strategies which reduce the amount of memory from 50% to 75% on real-world data sets and reduce the number of reading/writing operation from/to memory. These improvements open up the possibility to perform computation of distance-based network metrics on dynamic network at the fine-grained level of single edge addition/removal even on common hardware. Notice that for medium-sized network ($N \sim 10^5$ nodes) the amount of memory required to store NxN integer entries (1 byte each) is ∼10 GB which is close to the current limit of commodity hardware. The worst memory reduction of the SBGM approach experimentally found on real-world networks is about 50% (see Fig. 6.4) which reduces the previous number to 5GB, commonly available on standard hardware. The strategies proposed are based on the common underlying idea to drop unnecessary or redundant entries stored in the full geodesic matrix D. As a side result, the BGM approach provides a more efficient way to compute all-pairs shortest paths also on static networks. The larger problem of computing distances between all pairs of nodes in the graph is decomposed into the smaller problems of computing all-pairs distances within a single bi-component. Even if not all entries are explicitly computed the BGM still provides access to the value of distance between al pairs of nodes which can be used to construct important network metrics such as closeness centrality, global efficiency, information centrality.

Even though NetworkL has been primarily developed to test the SGM, BGM, and

SBGM approaches the code has reached a level of maturity which makes it suitable for computation in real-world scenarios. I have performed careful testing and validation by comparing more than 7×10^8 entries of the geodesic matrix produced by the library against those computed by Dijkstra algorithms implemented in NetworkX [134]. The future development roadmap includes the implementation of additional methods for the dynamic update of the BGM, the implementation of the combined SBGM approach, the manipulation of disconnected graphs, and the porting on C language. Contribution and feedback form developers and researchers are encouraged on the project page `http://github.com/networkl/`.

Chapter 7

Conclusions and future work

This thesis proposes a novel methodology for the study of innovation ecosystems and the prediction of various forms of success. From a theoretical perspective, I have conducted an extensive review of the most recent literature in the domains of complexity science, science of success, social and management science, and I have highlighted their link to the study of innovation processes. I have shown how the various theories and methodologies offered by these fields can be suitably integrated to provide a comprehensive and interdisciplinary approach to the study of innovation processes. In particular, I have outlined how the fundamental nature of innovation has changed over the last few decades, from a traditional form of technological innovation, characterised highly complex knowledge and specialised expertise, to a modern form of innovation stemming from recombinations of already existing technologies, involving less complex knowledge, and emerging from collective contributions of interacting individuals or firms. This radical shift prompts the need to revise the current understanding of innovation processes

and offers the opportunity to test old versus new hypotheses about the determinants of success in innovation practices. In my view, modern innovation is regarded as a socially-aided heuristic search process and a complex collective phenomena whose outcomes can not be explained simply in terms of individual agents' actions. It is indeed argued by many that the process of innovation cannot be well understood without paying attention to the social interactions among all actors involved (e.g., inventors, scientists, start-ups founders) [64, 164–166]. In this sense, social networks play a crucial role in the creation and diffusion of knowledge because they provide the structural foundations through which ideas can flow among individuals and can be integrated into novel recombinations [24]. For these reasons, in the section *"Innovation ecosystems through the network lens"* of Chapter 3, I have adopted approaches and methodologies borrowed from complexity science and social-network science to conceptualise potential indicators of success in innovation ecosystems. From the methodological point of view, these indicators have been operationalised in a set of network-based measures to characterise the various forms of access and exposure to knowledge. First, in Chapter 4, I have shown that the centrality score of a start-up firm in the WWS network is strongly correlated with the firm's long-term success. Remarkably, the method proposed is able to predict the exceptional success of the company WhatsApp only after 6 months since its foundation date and prior to the very first financial investment by Sequoia Capital. Results provide empirical support to the idea that networks of interaction between start-ups' members have a strong impact on the firm's performance and success. Second, in Chapter 5, I have investigated the extent to which an author's personal interdisciplinarity and exposure to others' knowledge impact upon the author's scientific performance. My results indicate

that scientists bear opportunity costs as they begin to diversify their background, at least until they become highly interdisciplinary. Moreover, scientists with groups of collaborators spanning many different areas are more successful than those with collaborators focused on one or few overlapping areas.

Key contributions.

The novel contributions in this thesis can be summarised as follows:

1. I have proposed a novel perspective for an interdisciplinary and comprehensive study of innovation processes in today's fast-changing society and economy. The main assumption on which this prospective rests is the idea that innovation lies within social relationships rather than in individual minds;

2. I have proposed a set of hypotheses based on this idea and tested them in two empirical domains: scientific production and the ecosystem of start-ups;

3. To account for sociality in scientific production, I have introduced the distinction between personal interdisciplinarity and social interdisciplinarity and have investigated empirically their relationship with an author's scientific performance. I have also studied the research strategies through which authors rely on their collaboration networks to amplify their personal interdisciplinarity over time. To my knowledge, this is the first study that complements research on interdisciplinarity by explicitly accounting for: (i) authors' time-varying collaboration networks; (ii) variations in authors' research strategies and performance over their entire careers; and (iii) various levels of granularity in the analysis (e.g., specialty, discipline).

4. To account for sociality in the start-up ecosystem, I have proposed and tested a

method for the prediction of long-term success of start-up companies based on formal affiliations and the network of professional connections. To my knowledge, my work offers the first study of the topological determinants of success of start-ups at a worldwide scale.

5. I have conducted an extensive data collection and cleaning process of the Crunchbase.com data resulting in the WWS network which will be soon made publicly available to the research community on the page `http://maths.qmul.ac.uk/~mbonaventura`;

6. To aid my empirical studies, I have implemented an incremental graph algorithm for the dynamic update of shortest paths. NetworkL is the first public python package which allows researchers and data scientists to re-compute shortest paths with incremental strategies. I have also proposed and formalised the concepts of sparse geodesic matrix (SGM) and sparse biconnected geodesic matrix (SBGM).

7.1 Implications for research and practice

My work has various implications for research. First, the new perspective proposed in this thesis can assist researchers in designing empirical investigations of innovation ecosystems. This new perspective arises from the combination of techniques and theories derived from complex network science and social sciences, and it offers several hypotheses about the determinants of success in modern innovation. Second, the empirical study of authors' interdisciplinarity has profound implications for current research in bibliometrics. Recent work on interdisciplinarity has focused mainly on the bene-

fits and disadvantages associated with authors' diversity of knowledge and background [39, 111, 167]. However, research on social capital and innovation has also suggested that performance is enhanced by the opportunities to gain and recombine knowledge offered by the network in which individuals are embedded [6, 54, 62, 63, 168]. In my study, I have integrated the individual and social perspectives, and proposed a conception of interdisciplinarity that extends beyond the boundaries of the scientists' background to also include their collaboration networks. I extended previous work by uncovering the competitive advantages of collaborative strategies for sustaining knowledge diffusion and acquisition. Third, my work on start-up firms is the first one in which innovation dynamics are observed at the fine-grained resolution of human interaction. It is also the first study to shed light on the impact of network structures on firms' performance at a global scale. Previous work has investigated how knowledge transfer impacts upon the performance of start-ups by using data on patents, inter-organisational collaborations, and co-location of firms to infer information flows and exchange [26, 96–100]. Other studies have analysed social networks (e.g., inventor collaboration networks) to unveil the microscopic level of interactions among individuals [169]; yet their scope has been limited mostly to specific industries or small geographic areas and observation periods [101, 102]. Owing to lack of data, the role of the global network that underpins knowledge exchange in the worldwide innovation ecosystem has been largely overlooked. Equally, the competitive advantage of differential information-rich network positions and their role in opening up, expediting, or obstructing pathways to firms' long-term success have been left largely unexplored. My contribution significantly overcomes these limitations and paves the way for more comprehensive studies of innovation processes.

Policy implications for scientific production. My empirical study on interdisciplinarity has far-reaching implications for research practice and policy. Opening up the black box of the scientist's knowledge to also account for collaboration networks paves the way for more integrated approaches to scientific production that borrow insights from bibliometrics and citation analysis, complex networks, cognitive science and the sociology of science. My findings can also inspire individual scientists to shape and sustain successful careers, research institutions to strengthen their scientific reputation and profile through effective recruitment policies and internal evaluations systems, and funding bodies to award research grants to projects with the highest potential impact.

Policy implication for start-up ecosystems.. In Her speech on the 18th May 2016, Queen Elizabeth II announced The Digital Economy Bill, later introduced in the House of Commons in July 2016. The Bill includes a range of measures to *"make the United Kingdom a world leader in the digital economy"* and to support new digital industries. Similarly, in 2010 the President of the United States Barack Obama launched the *Startup America* program, while the European Commission launched the *Startup Europe* program. Modern innovation is strongly driven by distributed efforts of thousands of digital innovators and start-ups. As an example, the digital tech industries are growing 32% faster than the rest of the UK economy, and those digital tech industries are creating employment opportunities accounting for 1.56M jobs across the UK. However, if on the one hand the net impact of start-ups on employment and wealth is positive, on the other hand the mortality rate of newly born creative businesses remains quite high. With a startup success rate of about 10%, the full potential of digital ecosystems is far from been unleashed, and there is still great room for improving the way innovation is managed and

monitored at the government's level. My work can help to build data-driven methodologies for informing policy decisions and maximising the potential of digital ecosystems. These methodologies are indeed crucial for governments that want to play a leading role in the digital market. The prediction method can help stakeholders devise and fine-tune a number of effective strategies, simply based on the underlying social network. Being able to estimate objectively the future potential of start-up companies will allow, on the one hand, investors to identify more quickly promising start-ups currently off the radar and, on the other hand, promising teams to stand out of the crowd, gain access to risk capital, and realise their potential more promptly.

7.2 Future work

The thesis suggests a number of new directions for future investigation. First, even though the main hypotheses proposed in Section 2.3 have been tested in two empirical domains, room is left for a more detailed study of each individual hypothesis. For instance, I have proposed that projects that have produced successful innovation are those which have gone through a fast-paced series of unsuccessful attempts. This concept suggests that, instead of measuring success directly, one could try to estimate the likelihood of future performance based on the number and characteristics of previous unsuccessful attempts. In practice, this could be tested by taking into account the number of unsuccessful companies founded by a person or the number of low-cited articles of a scientist. The study of scientists' interdisciplinarity can be extended in multiples ways. For instance, the role of global connectivity can be directly assessed by propos-

ing a different measure of interdisciplinarity that goes beyond the author's immediate local neighbourhood. The importance of accessing variegated pools of knowledge has been explicitly tested in chapter 5. A similar approach could be employed in the study of start-up ecosystems by decomposing the WWS network into a multilayer network in which the layers are the various professional roles. In this way one can study how access to various forms of information (e.g., know-how through employees, and business opportunities through mentors, advisors, and board members) impacts on performance. Additionally, one can investigate the extent to which other centrality measures, such as harmonic closeness, betweenness centrality, and PageRank, impact upon a start-up's long-term performance. It would also be worth investigating to which extent the method I have proposed to predict the success of start-ups can be used to predict economic performance of traditional businesses or the fluctuation in value of firms listed in the stock exchange markets. Lastly, the datasets I have collected are rich in other metadata which have not been fully exploited. A text-based analysis of article titles can prompt a different and more refined approach to the identification of the article topics. Moreover, data about skills, locations, and market sectors can further improve the study of start-up ecosystems. I have already undertaken some preliminary work in these directions in connection with the development of NetworkL.

Appendix A

Author's publications

Journal papers

1. Ciotti, V., Bonaventura, M., Nicosia, V., Panzarasa, P., Latora, V. (2016). Homophily and missing links in citation networks. *EPJ Data Science*, 5(1), 1.

2. P. Panzarasa, M. Bonaventura, (2015) The emergence of long-range correlations and bursty activity patterns in online communication, *Phys. Rev. E* 92, 062821.

3. M. Bonaventura, V. Nicosia, V. Latora (2014) Characteristic times of biased random walks on complex networks, *Phys. Rev. E* 89, 012803.

4. M. Bonaventura, V. Latora, V. Nicosia, P. Panzarasa, (2015) The advantages of interdisciplinarity in modern science, submitted to PNAS

5. Bonaventura, M., Ciotti, V., Panzarasa, P., Latora, V., Predicting success in the worldwide start-up network, (submitted).

Appendix B

Appendix of Chapter 4

B.1 Robustness and confounding factors

I have tested robustness of results against potential confounding factors such as location of the start-up, number of team members, and increase in the number of team members. Other factors such as the age of the founders, or age of the team members, can be in principle tested using the Crunchbase data set. However, since the date of birth is not one of the mandatory fields on the Crunchbase website, the disproportion between number of people reporting their date of birth and the ones that do not is very high (restricted to the US territory 292,684 people's records do not report age, and only 26,417 have age).

I have tested the performance of the prediction method by only selecting from the monthly open-deal lists from the start-ups based on a given location. Indeed, when the ranked list contains start-ups from any region of the world one can hypothesise that the top of the list would be dominated by firms which are based in notoriously prosperous

regions (e.g., Silicon Valley). In this case the higher concentration of successful start-ups at the top of the open-deal list could be a result of the location effect only and the centrality in the WWS network only reflects the location. I have assessed this problem directly by restricting the analysis to specific regions, including California (US). As I will show in a few paragraphs later, the results obtained at a worldwide scale are confirmed at the level of single regions. This indicates that start-ups that are better connected in the WWS network (i.e., that have higher network centrality) have an advantage over other start-ups that, being located in the same region, have potential access to the same set of opportunities (funding, business partners, access to talent, fiscal regulation, cost of running the business). The downside of splitting the open-deal lists by location is that the size of the list reduces significantly. For instance, in the UK, the number of start-ups belonging to the WWS network and included in the open-deal list prior to January 2004 is smaller than 20. In such cases analysing the success rate of the Top 20 is meaningless. As the sample size decreases the results became noisy. To improve the ratio signal/noise I have selected the top five regions per number of start-ups: the 3 US states of California (10,105 start-ups), New YorK (3,553), and Texas (1,140), United Kingdom (2,413), and Israel (592). For the California region it is still possible to analyse the success rate of the Top 100 while for all other regions I studied the evolution of the Top 10. All results are reported in Fig B.1.1. In almost all locations the success rate is higher than the random expectation for most of the observed period.

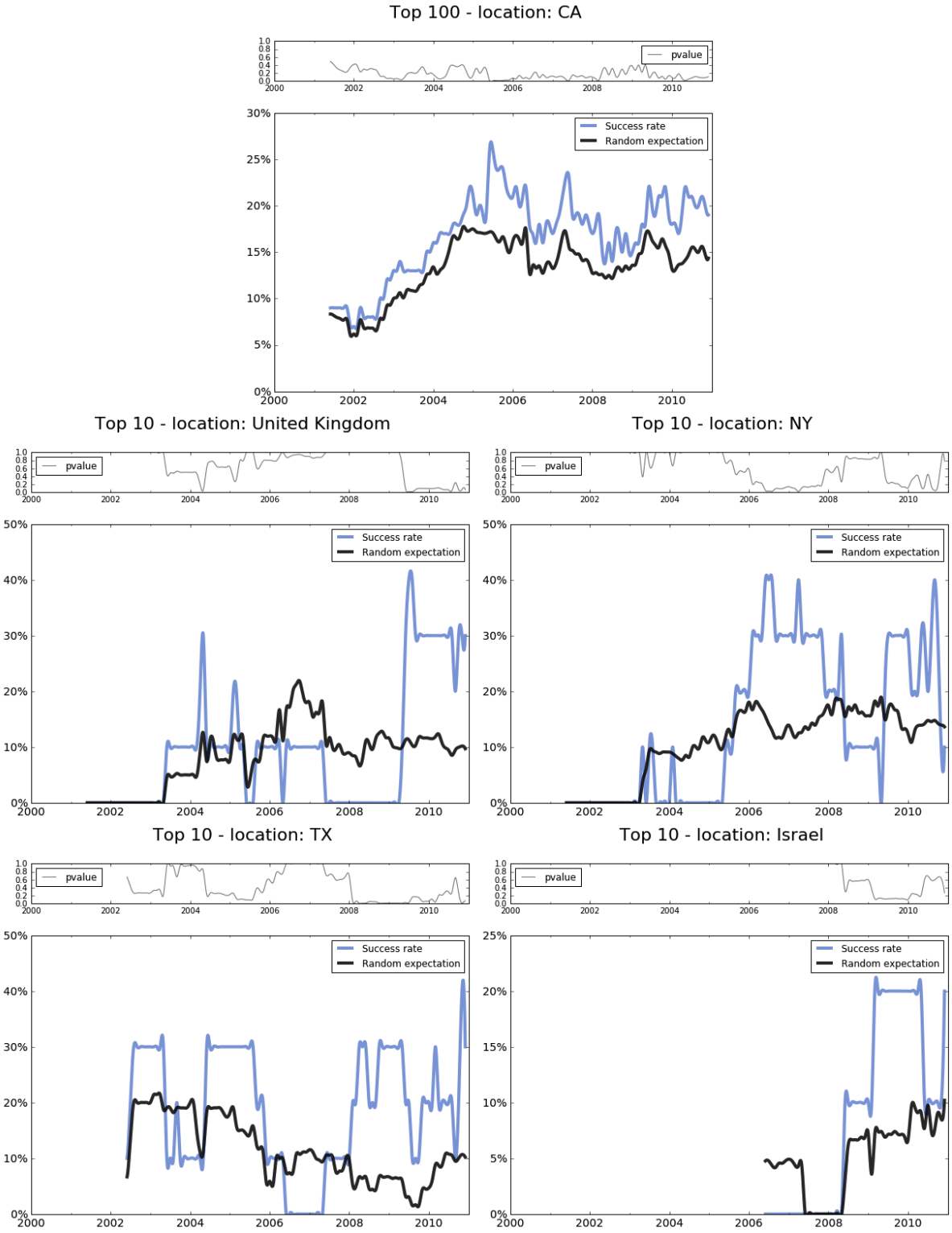

Figure B.1.1: Monthly success rate in the top five regions per number of start-ups.

In the case of the UK, the method performs badly up to 2009, after which the success rate raise to 3 times the random expectation and it is statistical significative (i.e., low p-values). It is worth noticing that 2010 is an important year for start-ups in general and for the ones based in London in particular. Indeed 2010 is the year of the launch of Obama's Startup America program and the year in which Google sets the basis for the London's start-up ecosystems by opening the Google Campus co-working space. In general the random expectation sits around the value $10\% - 15\%$ which is comparable with the industry standard of popular accelerator and investment firms. For example, the famous accelerator *500-Startups* has an overall success rate of 10% with 1,054 investments and only 120 companies acquired or publicly traded[1]. In this sense the results confirm that the proposed methodology can provide effective ways to improve investment practices. Moreover, the success rate for the California region is always greater than the random expectation and greater than the above mentioned industry standards. It is also worth noticing that the temporal trend is significantly different from the one obtained at the worldwide scale. While in Fig. 4.6(a) the peak occurs around mid-2003, in the California region the best performance is achieved in 2006 with a rate of 26% and it is followed by a stabilisation around 20%. This suggests that, while California is the leading region in terms of the number of start-ups, it's impact on the results at a global scale is limited and the plot in Fig 4.6(a) is not dominated by this leading region. Lastly, the plot reveals that the financial crisis had a small impact on the success rate in the California region, compared to other regions. In Fig. 4.6(a) the financial crisis drop occurs around year 2008 while in the same period California's trend seems stable around 20%. I have also computed the number of team members per

[1] source: Crunchbase.com

start-up in each month, their monthly increase, and studied the extent to which these indicators can be used to predict performance. Fig. B.1.2 shows the distribution of the maximum number of team members for all start-ups included in the open-deal lists. Since all firms have at least one team member when they join the lists, the figure implies also that the increase in the number of team members during the open-deal period is equal to zero for the large majority of the companies. Notice that the plot has a logarithmic scale and that the number of companies with more than 1 team member is only 596 across the whole observation period. Given the particular shape of the distribution it is very hard

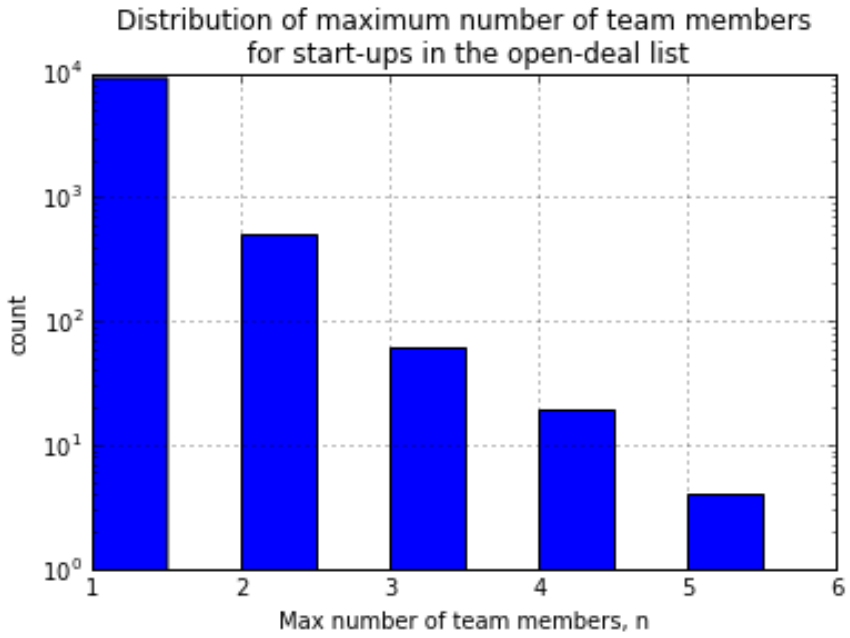

Figure B.1.2: Distribution of the maximum number of team members for all start-ups included in the open-deal list.

to obtain unambiguous ranking by using the number of people in the team as a success predictor (indeed nearly all start-ups will occupy the same position in the ranking as the majority of them have only 1 team member during the period in which they qualify to be part of the open-deal list). The analysis of groups of companies with an equal number

of team members has similar shortcomings. Indeed the group containing start-ups with 1 team member is likely to provide the same results as the one obtained by using the closeness centrality as a success predictor. Groups containing more then 1 team member have instead a negligible number of start-ups and are unlikely to provide any meaningful signal. Despite these various shortcomings, I have computed the success rate by using the monthly number of team members as a ranking metric. Fig. B.1.3 confirms that ranking based on the number of team members provides a success rate that is comparable with the one obtained by chance (the high variability in the success rate is due to the high number of companies with exactly one team member). An almost equal result is obtained for the monthly increase in number of team members. Fig. B.1.4 shows the success rate obtained for companies that have exactly one team member and confirms that the results are qualitatively similar to the one obtained for the complete open-deal list.

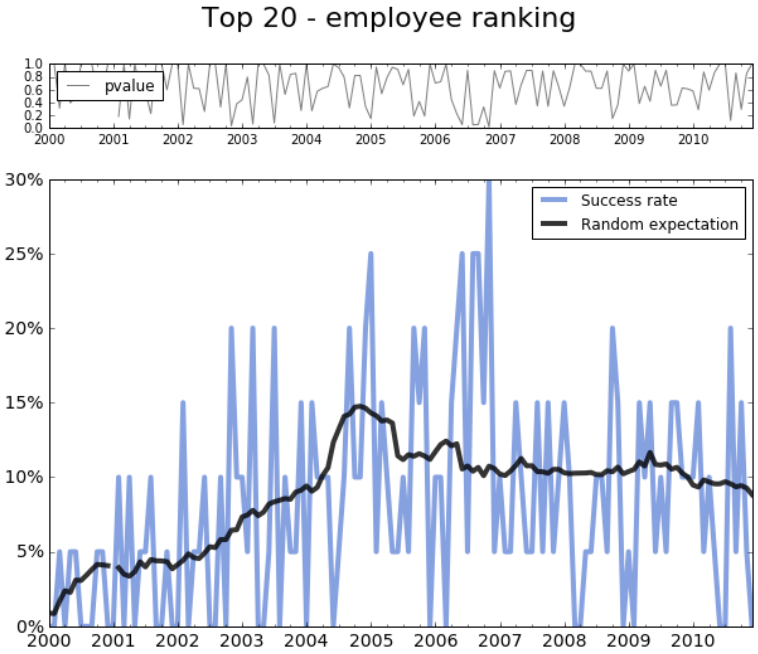

Figure B.1.3: Monthly success rate computed using the number of team members as ranking method.

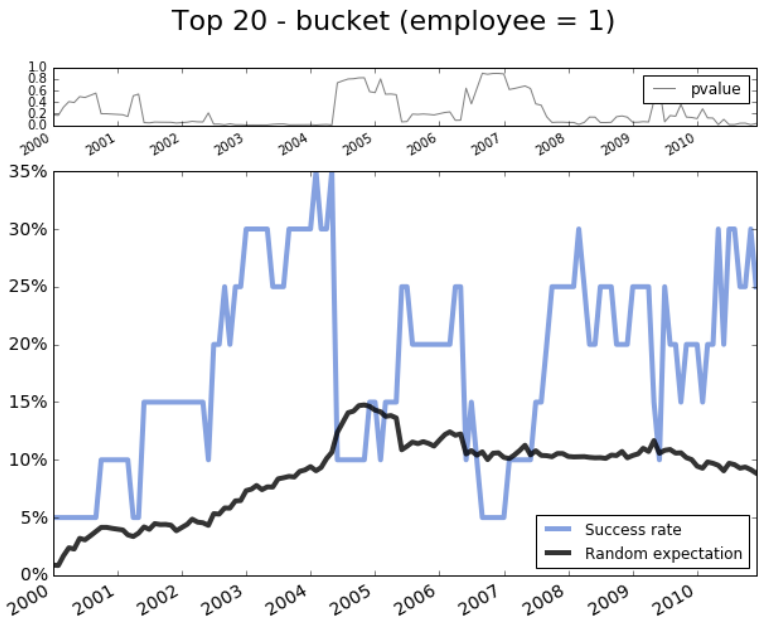

Figure B.1.4: Monthly success rate for companies with exactly one team member computed using the closeness centrality as ranking method.

To summarise, the results obtained by using closeness centrality as a ranking metric have proven to be robust against potential location biases or other confounding factors. While the number of team members is largely used in current venture capital practices to identify growth of more mature organisations, in the case of very early-stage start-ups the number of team members is hardly a distinguishing feature. Centrality measures in the WWS network have proven to be valuable indicators for investors that want to discover the best opportunities within the multitude of small innovative firms.

B.2 Fingerprints of start-up cities

The city is a convenient unit of analysis as it imposes a physical boundary to the local communities. Indeed, we expect to find a denser network of interactions within the same city (community) than across different cities. For simplicity we use the term *start-up ecosystem* to identify the subgraph whose nodes (companies and people) are associated with a certain city. I propose here a methodology to characterise various start-up ecosystems and to outline their differences and similarities in a quantitative way. In this analysis I have considered only the subset of cities which have at least 100 start-ups. First, for each ecosystem, I count the professional roles associated with all the links between a start-up and an individual. An example of the normalised distributions of roles in four US cities is reported in Figure B.2.5. Contrary to what one might

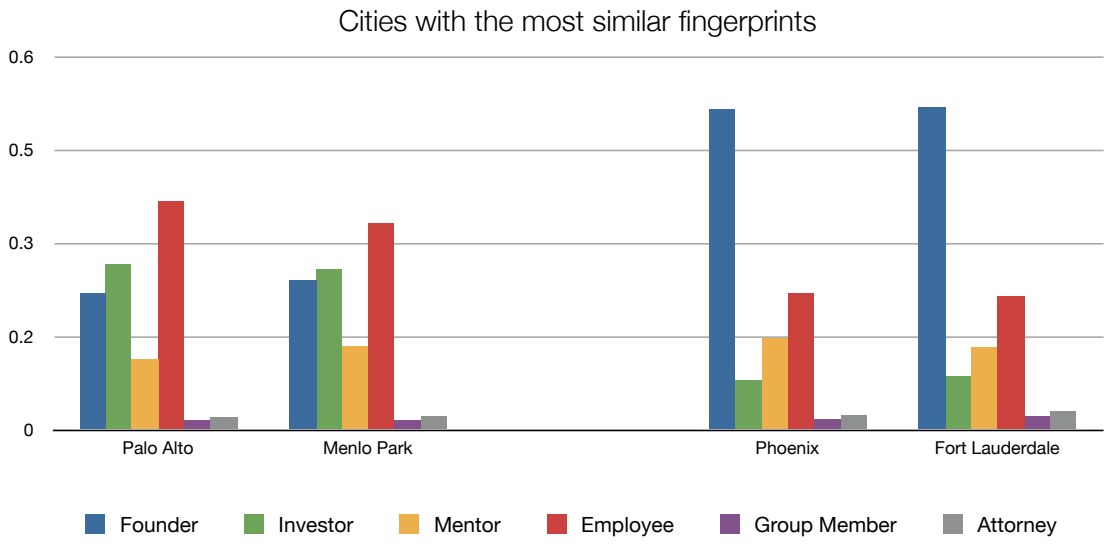

Figure B.2.5: Normalised distributions of roles in four US cities. Each histogram represents the *fingerprint* of the start-up ecosystem. The figure reports the two pairs of cities with the most similar fingerprints. The similarity is measured as Euclidean distance between two 6-dimensional vectors each associated with a city (see main text).

instinctively think, the histograms do not reflect the number of people with a certain role. Indeed the same person may have multiple roles in different start-ups, and all his/her roles contribute to the counting. In other words, we are considering the property *role* associated with each edge, rather than a property associated with the person. In this way we are able to associate an ecosystem, and its network, with a *fingerprint* which reflects its unique pattern of start-up activity. In some cities the founders' activity may be more pronounced (e.g., the blue pick of Fort Lauderdale and Phoenix) while other cities may display a more balanced distribution of roles (e.g., Palo Alto and Menlo Park).

The various fingerprints may reflect several underlying processes. For example, start-ups that receive substantial funding are more prone to quickly hire employees than others in which the founders, during the early stages of the company, take also the responsibility of project development (and do not consider themselves as employees). Some ecosystems, such as the one in the U.S., are more risk prone and display a greater fraction of investor roles then European ones. In order to study similarities and differences among the various ecosystems I computed the Euclidean distance between the 6-dimensional vectors (i.e., the normalised histograms) associated with each city. In particular, Figure B.2.5 shows the two pairs of cities with the most similar profiles. I then used the distance matrix between all pairs of cities to perform hierarchical clustering analysis. The clustering method used is *complete linkage*, and the associated dendrogram is shown in Figure B.2.6. Five main clusters are identified by the Elbow criterion. The fact that the majority of the clusters merge at a very small distance threshold (~ 0.05, i.e., bottom part of the dendrogram in Fig. B.2.6) indicates that cities within the same cluster are similar to each other, while there is a significant distance between the fingerprints of

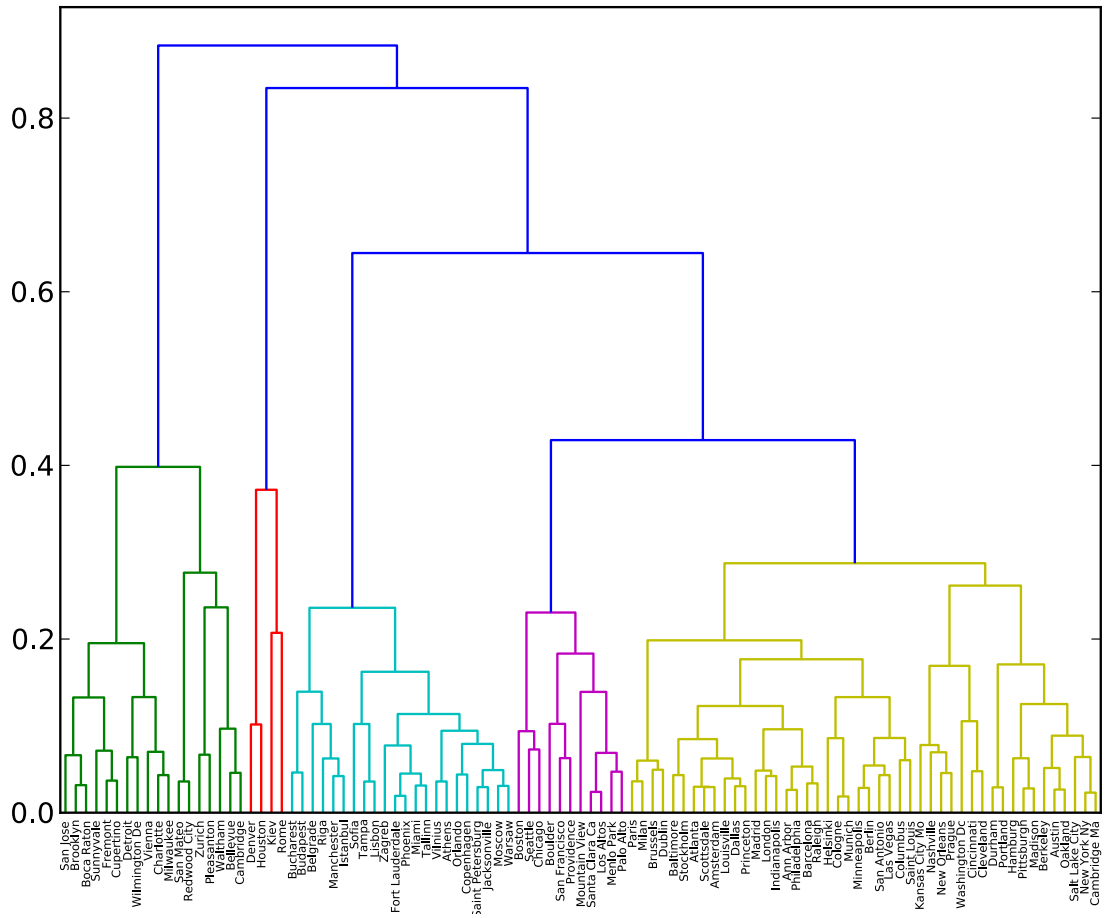

Figure B.2.6: Dendrogram resulting from the hierarchical clustering analysis.

cities belonging to different clusters.

In addition, I have found that some clusters have high degree of overlap with the geographic location (EU or US) of the cities included in the cluster. Figure B.2.7 shows the percentage of EU and US cities within each cluster and the representative fingerprint of each cluster, i.e., the 6-d vector constructed as the average (component by component) across all vectors belonging to the same cluster. The representative fingerprints reveal a distinct signature for European and US cities. While European cities are dominated by founders and mentors roles, the US ones are rich in employees and investors roles.

We conjecture that this result reflects the different attitude and approach of U.S. and EU citizens in the creation of new innovative activities. European citizens are prone to initiate new business activities, thus linking ownership and control (e.g., by founding a company) or to help others do so (e.g., by mentoring other companies). By contrast, U.S. people tend to focus more on either the financial aspect of the business or the execution stages, and their activities are therefore mainly devoted to providing the funding (investors) and to executing the business plan (employees).

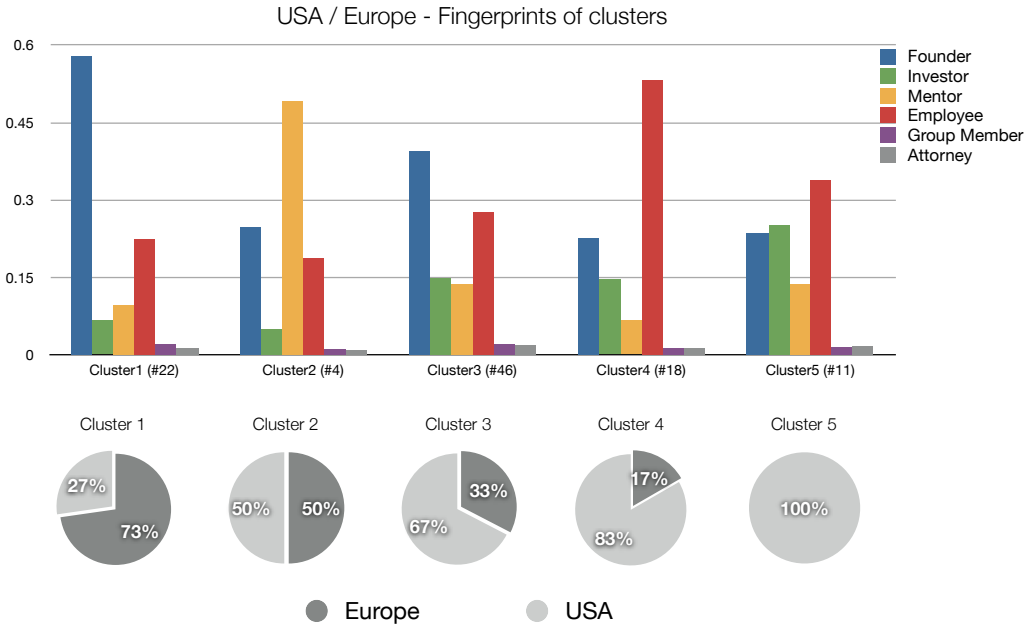

Figure B.2.7: The percentage of EU and U.S. cities within each cluster and the representative fingerprint of each cluster, constructed as the average across all 6-dimensional vectors belonging to the same cluster. The number in parenthesis (#X) indicates the number of cities within each cluster. Going from left to right, i.e., from EU-based to U.S.-based clusters, we notice a decrease of the fraction of founders roles and an increase of the fraction of investors and employees roles

I have also performed a similar analysis at a greater degree of detail by considering the

combination of all pairs of roles of the same person. An individual that holds two roles in two distinct companies acts as bridge between the two companies [80]. The interaction between the two companies can have a different nature and impact, depending on the specific roles of the bridging person. As already illustrated in Section 3.1, an employee moving from a company to another can mediate exchange of knowledge and know-how; a common investor can facilitate access to business opportunities and partnerships; a successful entrepreneur can speed growth up, and help newly born companies to avoid mistakes by acting as a mentor. The combinations of 6 professional roles give 21 different types of potential interactions between companies, labeled as follows: FF, FI, FM, FE, FG, FA, II, IM, IE, IG, IA, MM, ME, MG, MA, EE, EG, EA, GG, GA, AA. For instance, the FI interaction identifies the potential exchange of information between two companies mediated by a person which has acted as founder in the first and as investor in the second. For simplicity in this analysis the time ordering of the roles is disregarded, i.e., I do not distinguish if the founding event occurred before of after the investment one. Using the labels listed above it is possible to construct a 21-dimension fingerprint for each city. Figure B.2.8 shows three pairs of cities with the most similar fingerprints. Figure B.2.9 shows that the typical 21-d fingerprint of European cities has a predominance of FF links, indicating that the interaction between start-ups in the associated city subgraph is mainly mediated by founders. In agreement with the result of the 6-dimensions analysis, the same figure also confirms that investors play a crucial role for the interaction between start-ups in U.S. ecosystems. The analysis based on the 21-d fingerprints is not only technically more refined but has also a more meaningful interpretation than the 6 dimensions analysis. While the latter provides

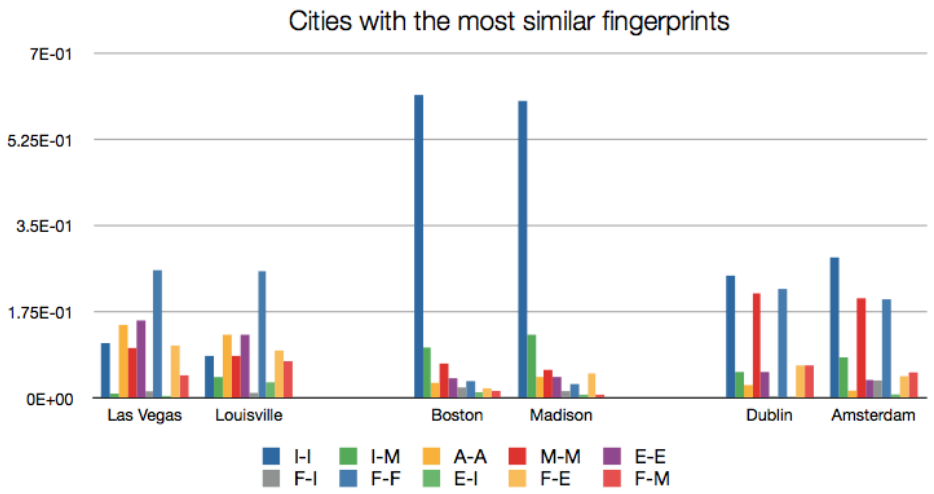

Figure B.2.8: Three pairs of cities with the most similar 21-d fingerprints.

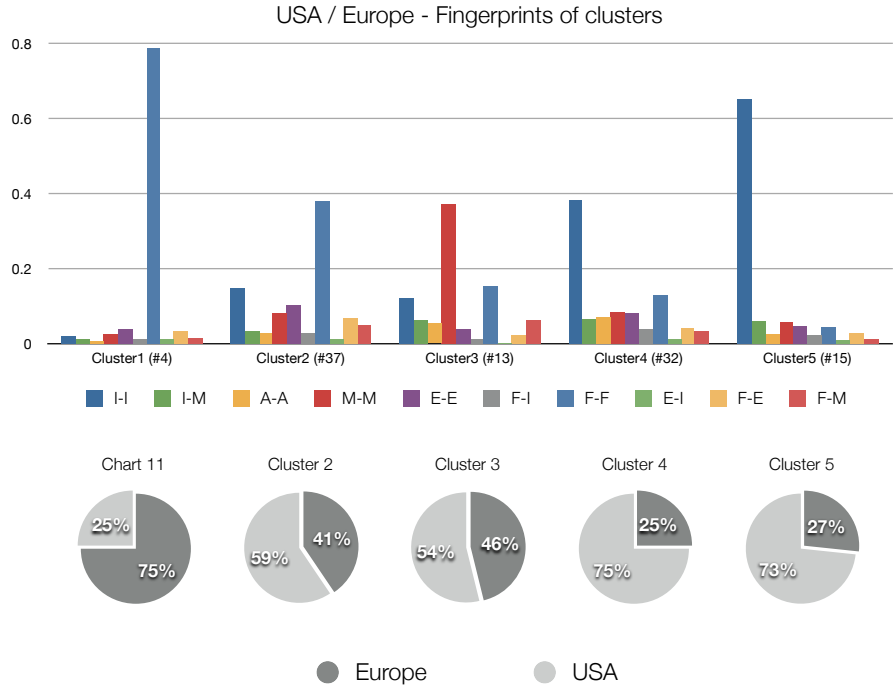

Figure B.2.9: The percentage of EU and U.S. cities within each cluster and the representative 21-d fingerprint of each cluster, constructed as the average across all 21-dimensional vectors belonging to the same cluster. The number in parenthesis (#X) indicates the number of cities within each cluster.

only information about the volume of the individuals' activity or propensity to take on certain roles, the former takes into account more deeply the structure of the network, the peculiarities of the links between companies and, in doing so, it captures the nature of exchanges between companies. Even though the 21-dimension analysis is limited to the local level of the network, it has the potential to shed light on the microscopic dynamics of interaction between companies, and to characterise with great detail the flow of ideas, knowledge, expertise, and opportunities as described in Chapter 3. Exploration of the interplay between the fingerprint of a city and the success of its startup ecosystem is left as future work.

Appendix C

Appendix of Chapter 5

C.1 Alternative measures of success

Average number of citations per article. I studied the relationship between interdisciplinarity and scientific performance by using the average number of citations per article over time as an alternative measure of success. The average number of citations of an author i at year t of his or her career is defined as:

$$\langle N^{cit} \rangle_i(t) = \frac{N_i^{cit}(t)}{N_i^a(t)}, \tag{C.1}$$

where $N_i^a(t)$ is the number of articles published by author i up to year t. The results from this additional analysis are reported in Fig.C.1.1. So constructed, this measure alleviate the differences in productivity of different authors and is more appropriate in the context of the second null model presented in section C.3.2.

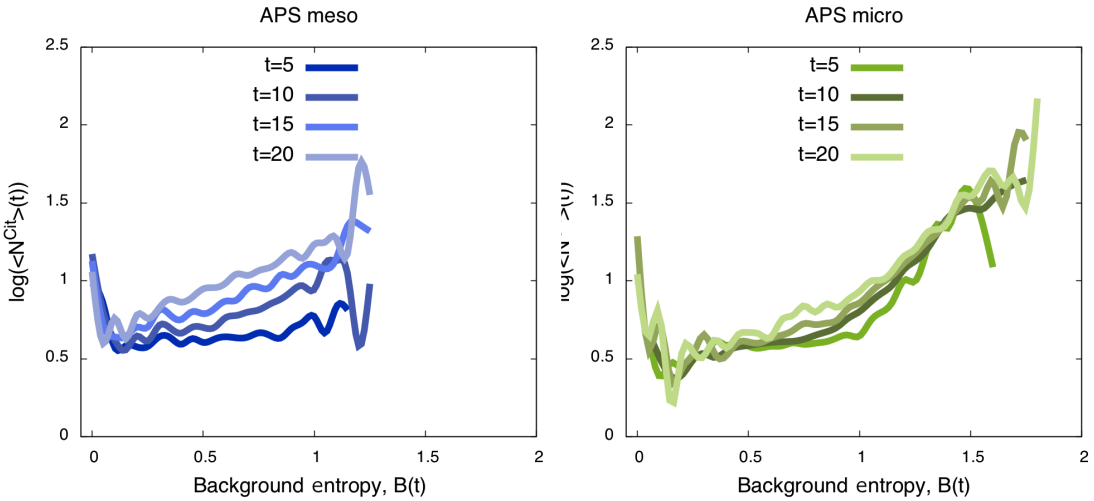

Figure C.1.1: Mean number of citations per article accrued by each author i, $\langle N^{cit}\rangle_i(t)$, averaged across all authors in the APS data set at different years t of their careers. The analysis is performed at two levels in the PACS code hierarchy: (i) the micro level (right-hand panel), which includes all 6 digits in the PACS codes, and (ii) the meso level (left-hand panel), which includes only the first 4 digits.

C.2 Disambiguation methods

In order to control for possible biases resulting from homonymous scientists, I replicated the analysis based on the APS data set employing three different strategies for name disambiguation. The first two strategies are variations of a standard method for initial-based name disambiguation, while the third one, adopted for producing all results in Chapter 5, relies upon a refined and comprehensive disambiguation method based on records about institutional affiliation, the collaboration network, and the citation network. In all strategies, first I removed all special characters from the records of each author's surname, given name, and middle name. I also filtered out surnames shorter than three characters to further reduce the risk of conflating two different homonymous authors within the same name. A given article was excluded entirely from the analysis

only if none of its co-authors complied with the chosen disambiguation criterion.

First name disambiguation strategy. The first disambiguation strategy considers only authors for whom both the full given name and the initial of the middle name are given. I then created each author's *stringID* by combining full surname and both initials (e.g., h.e.stanley). If the initial of either the given name or the middle name was missing, the author was filtered out and excluded from the analysis. The APS data set cleaned through this first strategy includes a total of $N_P = 302,104$ articles and $N_A = 96,516$ unique authors.

Second name disambiguation strategy. The second strategy requires at least the given name be included in each author's records. In this case, the middle name is optionally included in the author's *stringID*. The APS data set cleaned through this second name disambiguation strategy includes a total $N_P = 380,695$ articles and $N_A = 117,613$ unique authors. The results obtained with the first two disambiguation strategies are quantitatively similar to those obtained with the third disambiguation method presented in the following. The third strategy for name disambiguation, inspired by the one described in [38], directly relies on authors' institutional affiliations. Therefore, in what follows I shall first outline how I conducted a preliminary analysis of all affiliation records included in the entire APS data set.

C.2.1 Disambiguation of institutional affiliations

Institutional affiliations are included in the APS data set as free-text string records associated with one or more authors in each article. A condition of strict equality of affiliation strings is not suitable owing to the presence of small differences between strings

representing the same institution (e.g., punctuation, spaces, street addresses). First, I extracted the entire list of unique affiliation strings, and I filtered them by removing the common stop-words and common symbols listed in Table 3-A.

stop-words and symbols	“university” - “ of ” - “P.O. Box” “P.O.B.” - “ and ” - “ di ” “PO” - “P.O.”’ - “,” - “;” - “#”

Table 3-A: Stop-words and common symbols used for disambiguating institutional affiliations.

To check whether two affiliation strings represent the same institution, I performed a similarity test based on the Jaccard index. I split the strings into a list of words by using white spaces as separators, and computed the Jaccard index between the two sets. The APS data set contains $N_{aff} = 361,838$ unique affiliation records, which results into $N_{aff}^2 = 130.926.738.244$ computationally demanding pairwise comparisons between strings. The distribution of values of the Jaccard index for all affiliation pairs is shown in Fig. C.2.2.

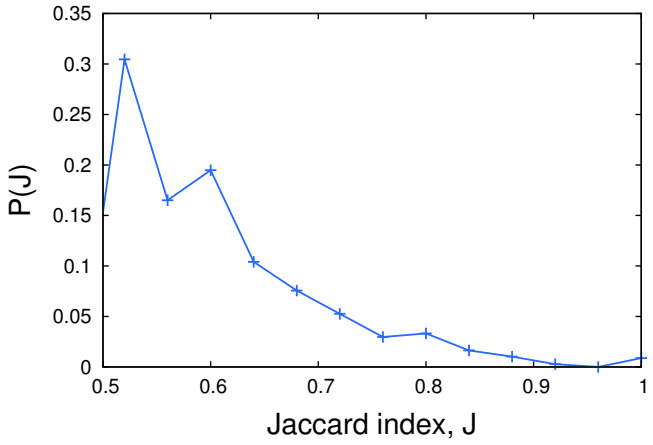

Figure C.2.2: Distribution of the values of the Jaccard index for all possible pairs of affiliation records. The threshold value was set at 0.6

Two affiliation records were conflated and assumed to be associated with the same insti-

tution if the Jaccard index was greater than a specific threshold value. As acceptance criterion, I chose a threshold value of 0.6. That is, two affiliation strings with a Jaccard index greater than 0.6 were conflated into the same institution, while two affiliation strings were associated with two distinct institutions if the index was lower than 0.6. There is no unique and objectively defined way for setting the threshold. In our case, the value of 0.6 was chosen based on: (i) the total number of unique institutions resulting from the disambiguation process, and (ii) a number of tests manually performed on various subsets of affiliations. As an example, during the disambiguation process the following four different strings were identified as similar, and thus referring to the same institution:

(a) "Department of Applied Mathematics, Queen Mary College, London E1 4NS, United Kingdom"

(b) "Department of Applied Mathematics, Queen Mary College, University of London, London E1 4NS, England"

(c) "Department of Applied Mathematics, Queen Mary College, University of London, London E1 4NS, United Kingdom"

(d) "Department of Applied Mathematics, Queen Mary College, University of London, London, England"

C.2.2 Third name disambiguation strategy

The previous criterion for disambiguating affiliations is fed into a third, more conservative method for the disambiguation of authors' names [38]. To speed up the computation, first I grouped authors with equal surnames (as this field is always included in the data) into distinct subsets, and I kept record of the articles they published. Then I checked all possible pairs of articles within each subset. In so doing, only authors with equal surnames were compared in the following way. For each pair of articles, and for each author i in article a_1 and author j in article a_2, I checked whether author i and author j passed a number of similarity tests. Specifically, I checked if the given names and middle names, where available, were compatible. If the full given name or the middle name was available, the comparison was performed by taking into account the entire string and potential misspellings. In particular, two string names were assumed to be compatible if they had a Hamming distance smaller than two. If given names and middle names were not compatible, then the corresponding authors were assigned to two distinct numeric IDs. On the contrary, authors were merged into a single identity if given names (and middle names) were compatible, and if at least one of the following conditions applied [38]:

- the corresponding institutional affiliations are similar (as outlined in the previous subsection C.2.1);
- the two articles share at least one similar co-author (string comparison);
- there is a citation between the two articles.

C.3 Null models

C.3.1 Reshuffling PACS codes and research categories

I compared the results shown in Fig. 5.1 and Fig. 5.2 with those obtained through a null model in which PACS codes (in the APS data set) and research category codes (in the WOS data set) were randomly shuffled. Codes were shuffled only across articles published in the same year. This choice enabled us to construct a null model that properly takes into account variations in interest and popularity of some disciplines as well as the appearance of new scientific fields over time. Once PACS codes or research categories are shuffled across articles, the newly computed values of background and social entropy do not reflect any longer an author's inclination towards specialised or interdisciplinary research. By contrast, the background entropy computed on a randomly reshuffled list of codes is expected to correlate only with the size of such list, i.e., with the number of codes, and, in turn, with the number of articles an author has published. I averaged the results (namely, the average number of citations obtained by authors with various values of entropy) over $1,000$ independent realisations of the null model, and I compared the results with the ones obtained with the real data. In the top panels of Fig. C.3.3 I plot the relation between background entropy and citations in the two data sets and in the null model. It is interesting to notice that the null model fails to reproduce the U-shaped relation between interdisciplinarity and success, since it predicts a monotonically increasing dependence of success on background entropy. Notice that the minimum values of background entropy obtained in the realisations of the null model are consistently higher than the minimum values observed in the two data sets, in

agreement with the fact that the random assignment of PACS codes and categories does not allow authors to remain focused on a topic. Moreover, the predicted average number of citations per author in the null model is consistently smaller than that observed in the data, especially in APS where the difference is of about one order of magnitude. In the bottom panels of Fig. C.3.3 I also report the relation between social entropy and success in the data and in the null model. It is also important to note that, by reshuffling PACS codes and research categories, this null model destroys the existing patterns of authors interdisciplinarity. In Fig. C.3.4 I show the distributions of author background entropy in the APS data set (top panels) and in the null model (bottom panels) for authors with different numbers of published articles (respectively, 5, 10, 20, 50, and more than 70, proceeding from the leftmost to the rightmost panel in each row). It is evident that the typical distribution of background entropy for authors with a certain number of articles is relatively wide and negatively skewed. Conversely, for the same number of published articles, the null model grossly overestimates the interdisciplinarity of each author, and the corresponding distribution of background entropy is symmetric and tightly peaked around a typical value.

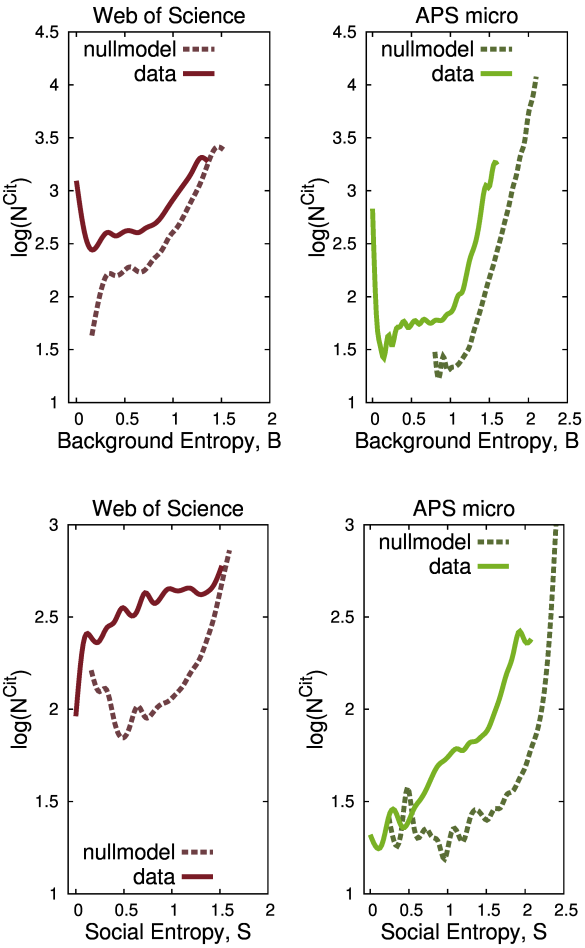

Figure C.3.3: Average number of citations in real data and in the first null model as a function of author background entropy (top panels) and author social entropy (bottom panels). The relation between background entropy and citation is monotonically increasing in the null model, while in the data set I observe a U-shape. Moreover, the random distribution of PACS and category codes forbids small values of background entropy, which are instead observed both in the APS and in the WoS data set.

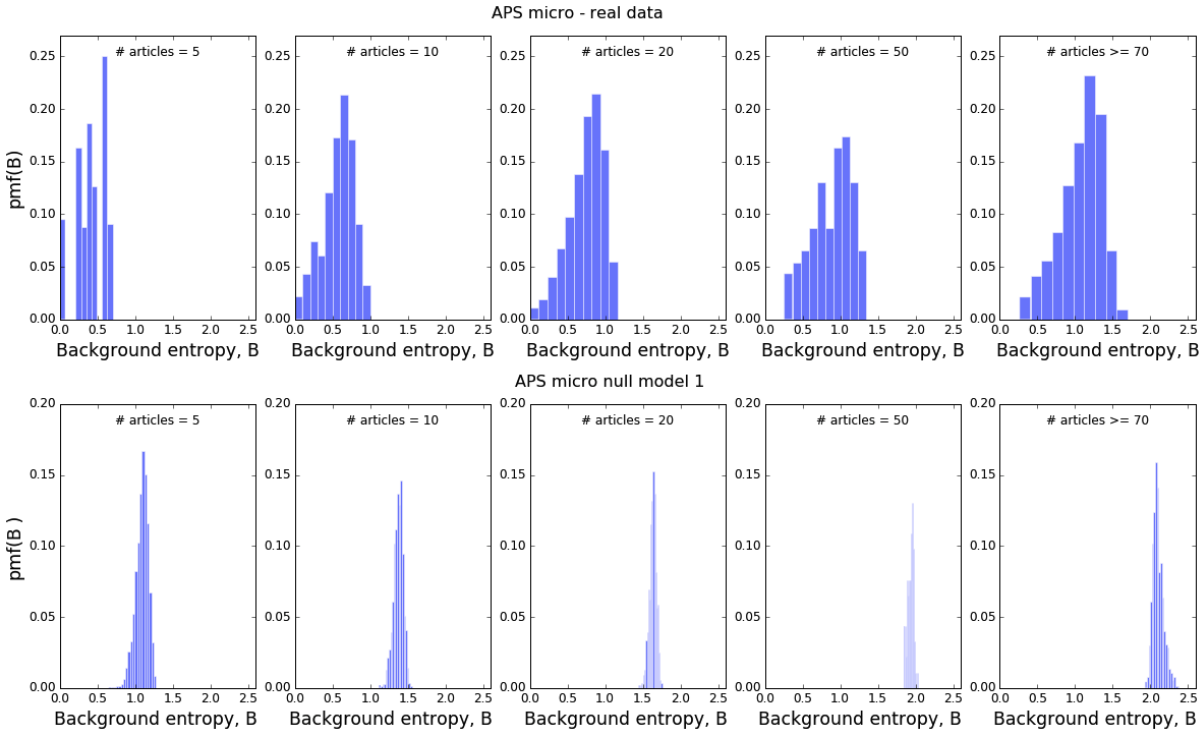

Figure C.3.4: Distributions of personal entropy in the APS data set (top panels) and in the first null model (bottom panels) for authors having published, 5, 10, 20, 50, and 70, respectively from left to right. The distributions of personal entropy observed in the real data set are very different from those expected if the PACS were chosen at random, and in particular are less peaked and centred around smaller values of background entropy.

C.3.2 Reshuffling the citation graph

I considered a null model aiming at assessing whether the U-shaped dependence between success and interdisciplinarity, shown in Fig. 5.1 and Fig. 5.2 of the main text, was simply due to the heterogeneity in the number of articles published by each author and to the differences in the length of their careers. This null model is constructed by reshuffling the network of citations among articles, while keeping fixed the association of authors and research categories to articles, and the length of the reference list of each article (namely, the out-degree of each node in the citation graph). In particular, each of the outgoing links of article a was randomly reassigned to one of the articles published before a. We notice that by doing so the number of citations received by a paper depends only on the age of the paper, and the total number of citations accrued by an author depends heavily on the number of published articles. However, in this null model the value of background entropy of each author (i.e., the x-coordinate of the author in the entropy/success plane) is identical to that observed in the data sets, since the associations of authors and PACS codes to articles is preserved. The comparison between the data (blue line) and the null model (green line) is reported in Fig. C.3.5, where I used as a measure of success of author i the average number of citations received by each article authored by i given by Eq. (C.1). It is evident from the figure that the null model is not able to reproduce the U-shaped relation between interdisciplinarity and success, since it predicts that the average number of citations per article would remain practically constant over the whole range of background entropy values.

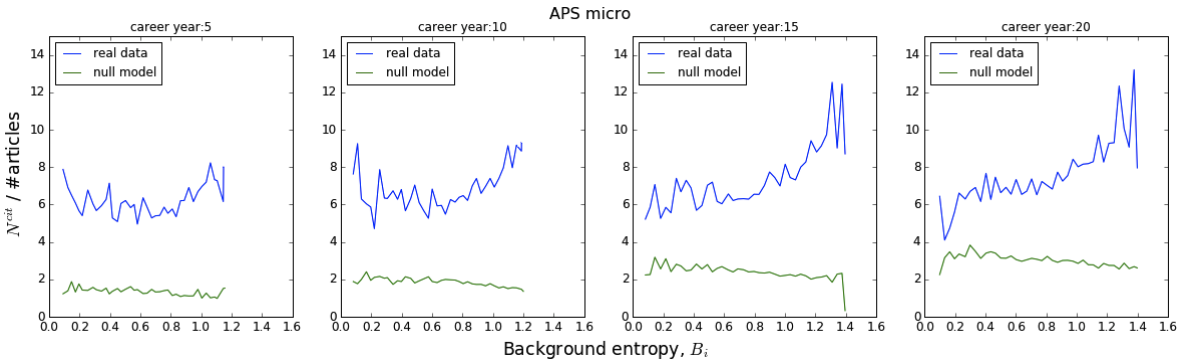

Figure C.3.5: Comparison between the average number of citations per article for each author in the APS data set (blue line), and the corresponding value observed in the null model where the citation network among articles is reshuffled but the background entropy of each author is kept constant (green line). It is evident that this null model is not able to reproduce the U-shaped relation between interdisciplinarity and success, allowing us to rule out the possibility that the U-shape is simply due to the heterogeneity in the number of articles published by each author and in the career length of different authors.

C.4 Distribution of citations within different ranges of authors' background entropy

In this section, I study the frequency distributions of the number of citations received by all authors falling within the same bins of background entropy. It is evident that each of the probability mass functions shown in Fig. C.4.6 is characterised by a peak around a typical value of citations. Contrary to the number of citations per author in the whole dataset, which have a power-law distribution, the distributions of citations restricted to given bins of background entropy are not broad. Indeed the presence of a peak in the frequency distribution indicates that a group of authors characterised by the same value of interdisciplinarity have a mean number of citations that is representative of the entire group. For instance, the distributions of citations accrued by authors with values of background entropy B lying within the ranges $[0, 0.05]$ and $[1.2, 1.25]$ (i.e., highly focused and highly interdisciplinary authors, respectively corresponding to the red line and yellow line in Fig. C.4.6) are both peaked around a value of $\log(N^{cit}) \sim 3$, while the distributions of citations for authors with background entropy ranging within $[0.3, 0.35]$, $[0.6, 0.65]$, and $[0.9, 0.95]$ (i.e., ranges of values that correspond to the local minima of the U-shaped curve) are all peaked around lower values of citations ($\log(N^{cit}) \sim 1.0$ in APS, and $\log(N^{cit}) \sim 2.25$ in WOS).

I have also assessed the statistical significance of the differences between these distributions of citations by performing the Kolmogorov-Smirnov (KS) test between each pair of distributions obtained within different bins. Fig. C.4.7 shows which pairs of distributions satisfy the KS test at the $p = 0.01$ significance level. Findings suggest that

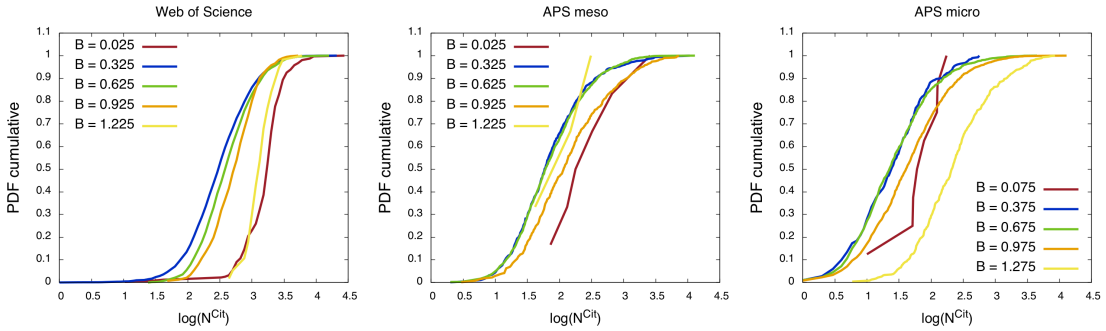

Figure C.4.6: Cumulative distributions of the number of citations received by authors within the same bins of background entropy

pairs of groups of authors associated with ranges of background entropy corresponding one to a local maximum and the other to the minimum of the U-shaped curve were sampled from populations with different frequency distributions of citations. By contrast, pairs of distributions corresponding to the two local maxima of the U-shaped curve are statistically significantly similar.

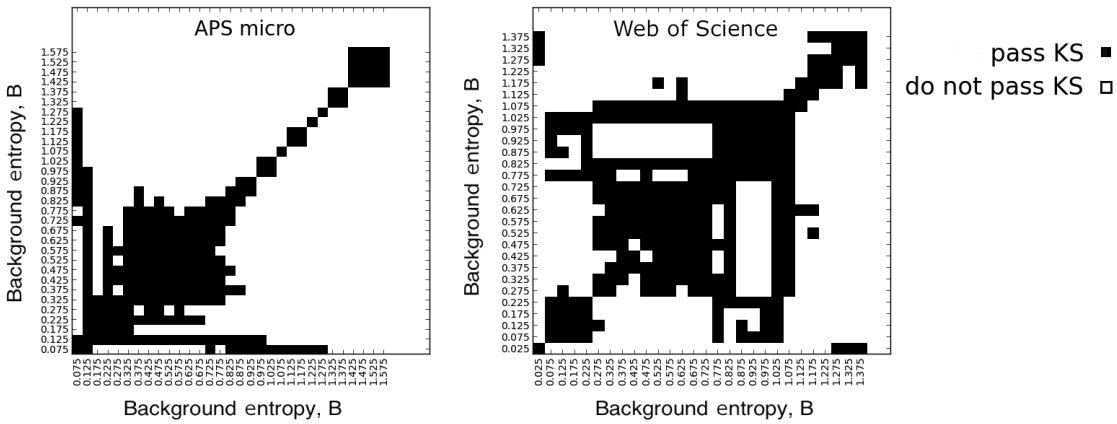

Figure C.4.7: Kolmogorov-Smirnov (KS) test for all pairs of cumulative frequency distributions of citations obtained by authors characterised by values of background entropy lying within different bins. Each bin is identified by the value of background entropy at the centre of the bin (e.g., the value 0.025 on the x-axis indicates the bin $B \in [0, 0.05]$). Each point in the matrix is a pair of distributions of citations shown in Fig. C.4.6. Black squares indicate pairs of distributions that pass the KS test, while white squares indicate pairs that do not pass the test.

In Fig. C.4.8 I show the distributions of citations centred around the mean for all bins of background entropy. I have fitted each of these distributions against a power-law, a lognormal, and an exponential distribution and assessed the goodness of the fit. The goodness of the fit is computed as loglikelihood ratio R between the two candidate distributions [170] and its statistical significance is provided by the associated p-values[1]. All pairs of R and p-values are reported in Fig. C.4.9. The negative values of R indicate that the second distribution (exponential) fits the data best than the power-law one. Instead, in the case of the lognormal distribution, the p-values are high indicating that none of the two distributions provide a better description of the data than the other. Overall this suggests that the exponential distribution is a valid description of the data and that, consequently, the contribution of "extreme" authors to the mean values of the U-shape is negligible.

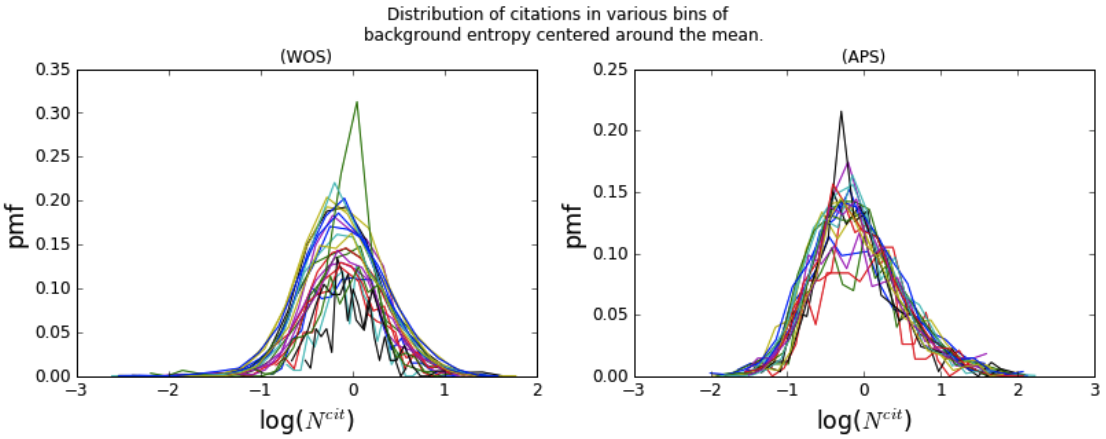

Figure C.4.8: Distributions of citations in various bins of background entropy centred around the mean. The various distributions exponential shape also confirmed by the statistical test.

[1]loglikelihood and pvalues are computed using the package *powerlaw* presented in [171]

APS

loglikelihood ratio	pvalue
-1.0154	0.0031
-3.7615	0.0004
-1.0467	0.1357
-0.2140	0.3829
-3.6951	0.0024
-2.3452	0.0007
-0.6429	0.2212
-1.6419	0.0780
-1.9739	0.0114
-0.2485	0.5569
-1.3859	0.1570
-1.5390	0.0365
-2.5554	0.0370
-2.1831	0.0909
-1.2654	0.2394
-21.3457	0.0000
0.0951	0.7779
-1.9943	0.0722
-7.3754	0.0000
-1.7232	0.0458

WOS

loglikelihood ratio	pvalue
-1.3191	0.0235
-0.7200	0.0285
-0.0412	0.8519
0.0036	0.9742
-0.3524	0.0239
-0.0149	0.8786
0.1482	0.7685
-0.4366	0.4676
-3.2369	0.0004
-0.7934	0.0277
-0.8654	0.1483
-5.2931	0.0000
-2.3832	0.0012
0.0984	0.7187
-1.9528	0.0440
0.0039	0.9335
-0.8864	0.1567
0.3040	0.5505
-1.5303	0.0025
0.0298	0.6623

Figure C.4.9: Loglikelihood ratios R and p-values for an exponential and power-law fit of the citations distributions in figure C.4.8. Negative values of R and small pvalues indicate that the exponential distribution should be preferred to the power-law one in most cases.

C.5 Historical trend

To asses whether the success of interdisciplinary studies is modern phenomenon only or if authors have benefited from interdisciplinarity also in the past, I have repeated the analyses by dividing the sample in two groups: authors whose career has ended before the year 2000, and those who have started their career after the year 2000. Notably the U-shaped trend is confirmed in both groups despite the advantage of interdisciplinarity and specialisation seem more pronounced in the most recent group (i.e., the number of rescaled citations for interdisciplinary and specialised authors is higher in more recent years).

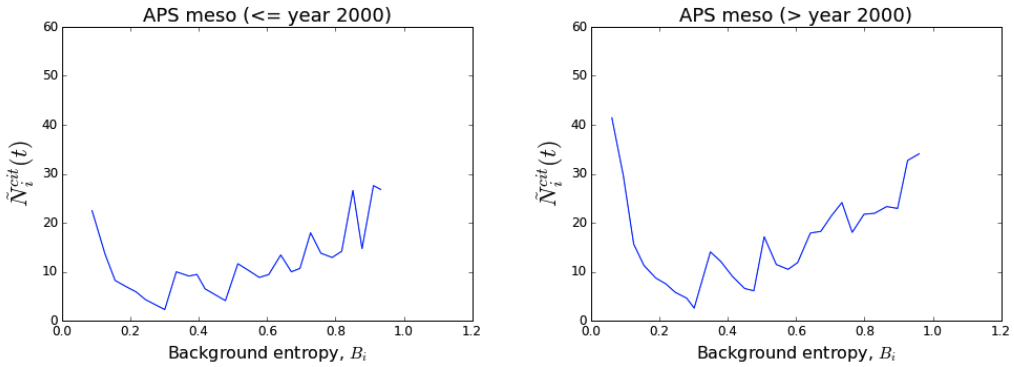

Figure C.5.10: Relation between citations and background entropy in two non-overlapping historical periods (before and after year 2000).

References

[1] Sawhney M, Wolcott RC, Arroniz I (2006) The 12 different ways for companies to innovate. *MIT Sloan Management Review* 47:75.

[2] Birkinshaw J, Bouquet C, Barsoux J (2011) The 5 myths of innovation: top 10 lessons on the new business of innovation. *MIT Sloan Management Review* 52:1–8.

[3] Lenski G (1979) Directions and continuities in societal growth. *Societal Growth. Free Press, New York* pp 5–18.

[4] Mokyr J (2002) *The gifts of Athena: Historical origins of the knowledge economy* (Princeton University Press).

[5] Fleming L (2001) Recombinant uncertainty in technological search. *Management science* 47:117–132.

[6] Sarigöl E, Pfitzner R, Scholtes I, Garas A, Schweitzer F (2014) Predicting scientific success based on coauthorship networks. *EPJ Data Science* 3:1.

[7] Sampson RC (2007) R&d alliances and firm performance: The impact of technological diversity and alliance organization on innovation. *Academy of Management Journal* 50:364–386.

[8] Radicchi F (2012) Universality, limits and predictability of gold-medal performances at the olympic games. *PloS one* 7:e40335.

[9] Petersen AM, Jung WS, Stanley HE (2008) On the distribution of career longevity and the evolution of home-run prowess in professional baseball. *EPL (Europhysics Letters)* 83:50010.

[10] Cintia P, Pappalardo L, Pedreschi D (2013) *"Engine Matters": A First Large Scale Data Driven Study on Cyclists' Performance* (IEEE), pp 147–153.

[11] Cintia P, Giannotti F, Pappalardo L, Pedreschi D, Malvaldi M (2015) *The harsh rule of the goals: data-driven performance indicators for football teams* (IEEE), pp 1–10.

[12] Heuer A, Mueller C, Rubner O (2010) Soccer: Is scoring goals a predictable poissonian process? *EPL (Europhysics Letters)* 89:38007.

[13] Ben-Naim E, Redner S, Vazquez F (2007) Scaling in tournaments. *EPL (Europhysics Letters)* 77:30005.

[14] Onody RN, de Castro PA (2004) Complex network study of brazilian soccer players. *Physical Review E* 70:037103.

[15] Mazloumian A (2012) Predicting scholars' scientific impact. *PloS one* 7:e49246.

[16] Penner O, Pan RK, Petersen AM, Kaski K, Fortunato S (2013) On the predictability of future impact in science. *Scientific reports* 3.

[17] Petersen AM, Jung WS, Yang JS, Stanley HE (2011) Quantitative and empirical demonstration of the matthew effect in a study of career longevity. *Proceedings of the National Academy of Sciences* 108:18–23.

[18] Uzzi B, Mukherjee S, Stringer M, Jones B (2013) Atypical combinations and scientific impact. *Science* 342:468–472.

[19] Wang D, Song C, Barabási AL (2013) Quantifying long-term scientific impact. *Science* 342:127–132.

[20] Servia-Rodríguez S, Noulas A, Mascolo C, Fernández-Vilas A, Díaz-Redondo RP (2015) The evolution of your success lies at the centre of your co-authorship network. *PloS one* 10:e0114302.

[21] Ma A, Mondragón RJ, Latora V (2015) Anatomy of funded research in science. *Proceedings of the National Academy of Sciences* 112:14760–14765.

[22] Pentland A (2012) The new science of building great teams. *Harvard Business Review* 90:60–69.

[23] Brown C, et al. (2014) *The architecture of innovation: Tracking face-to-face interactions with ubicomp technologies* (ACM), pp 811–822.

[24] Bettencourt LM, Lobo J, Helbing D, Kühnert C, West GB (2007) Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the national academy of sciences* 104:7301–7306.

[25] Inoue H, Liu YY (2015) Revealing the intricate effect of collaboration on innovation. *PloS one* 10:e0121973.

[26] Guzman J, Stern S (2015) Where is silicon valley? *Science* 347:606–609.

[27] Hale SA, Margetts H, Yasseri T (2013) *Petition growth and success rates on the UK No. 10 Downing Street website* (ACM), pp 132–138.

[28] Colombo MG, Franzoni C, Rossi-Lamastra C (2015) Cash from the crowd. *Science* 348:1201–1202.

[29] Salganik MJ, Dodds PS, Watts DJ (2006) Experimental study of inequality and unpredictability in an artificial cultural market. *science* 311:854–856.

[30] Weng L, Menczer F, Ahn YY (2014) Predicting successful memes using network and community structure. *arXiv preprint arXiv:1403.6199*.

[31] Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) *Everyone's an influencer: quantifying influence on twitter* (ACM), pp 65–74.

[32] Radicchi F (2011) Who is the best player ever? a complex network analysis of the history of professional tennis. *PloS one* 6:e17249.

[33] Yucesoy B, Barabási AL (2015) Untangling performance from success. *arXiv preprint arXiv:1512.00894*.

[34] Muchnik L, Aral S, Taylor SJ (2013) Social influence bias: A randomized experiment. *Science* 341:647–651.

[35] Aral S, Walker D (2011) Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science* 57:1623–1639.

[36] Aral S, Walker D (2012) Identifying influential and susceptible members of social networks. *Science* 337:337–341.

[37] Aral S, Muchnik L, Sundararajan A (2013) Engineering social contagions: Optimal network seeding in the presence of homophily. *Network Science* 1:125–153.

[38] Deville P, et al. (2014) Career on the move: Geography, stratification, and scientific impact. *Scientific Reports* 4:4770.

[39] Wuchty S, Jones BF, Uzzi B (2007) The increasing dominance of teams in production of knowledge. *Science* 316:1036–1039.

[40] Freeman LC (1977) A set of measures of centrality based on betweenness. *Sociometry* pp 35–41.

[41] Erdős P, Hajnal A (1966) On chromatic number of graphs and set-systems. *Acta Mathematica Hungarica* 17:61–99.

[42] Seidman SB (1983) Network structure and minimum degree. *Social networks* 5:269–287.

[43] Wasserman S, Faust K (1994) *Social network analysis: Methods and applications* (Cambridge university press) Vol. 8.

[44] Petersen AM, et al. (2014) Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences* 111:15316–15321.

[45] Ciotti V, Bonaventura M, Nicosia V, Panzarasa P, Latora V (2016) Homophily and missing links in citation networks. *EPJ Data Science* 5:1.

[46] Catalini C, Lacetera N, Oettl A (2015) The incidence and role of negative citations in science. *Proceedings of the National Academy of Sciences* 112:13823–13826.

[47] Saxenian A (1996) *Regional advantage* (Harvard University Press).

[48] Hwang VW, Horowitt G (2012) *The Rainforest: The secret to building the next Silicon Valley* (Regenwald Los Altos Hills, California).

[49] Lazer D, et al. (2009) Life in the network: the coming age of computational social science. *Science (New York, NY)* 323:721.

[50] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. *Physics reports* 424:175–308.

[51] Pontin J (2012) Why we can't solve big problems. *Technology Review* 115:26–31.

[52] Clark A, Chalmers D (1998) The extended mind. *analysis* 58:7–19.

[53] Argote L, Ingram P (2000) Knowledge transfer: A basis for competitive advantage in firms. *Organizational behavior and human decision processes* 82:150–169.

[54] Rodan S, Galunic C (2004) More than network structure: how knowledge heterogeneity influences managerial performance and innovativeness. *Strategic Management Journal* 25:541–562.

[55] Perry-Smith JE (2006) Social yet creative: The role of social relationships in facilitating individual creativity. *Academy of Management journal* 49:85–101.

[56] Fleming L, Sorenson O (2002) Navigating the technology landscape of innovation. *MIT Sloan Management Review* 44:15–24.

[57] Fleming L, Sorenson O (2004) Science as a map in technological search. *Strategic Management Journal* 25:909–928.

[58] Sorenson O, Singh J (2007) Science, social networks and spillovers. *Industry and Innovation* 14:219–238.

[59] Stuart TE, Sorenson O (2005) in *Handbook of entrepreneurship research* (Springer), pp 233–252.

[60] Fleming L, Mingo S, Chen D (2007) Collaborative brokerage, generative creativity, and creative success. *Administrative science quarterly* 52:443–475.

[61] Coleman J, Katz E, Menzel H (1957) The diffusion of an innovation among physicians. *Sociometry* 20:253–270.

[62] Burt RS (2004) Structural holes and good ideas. *American journal of sociology* 110:349–399.

[63] Sosa ME (2011) Where do creative interactions come from? the role of tie content and social networks. *Organization Science* 22:1–21.

[64] Powell WW, Koput KW, Smith-Doerr L (1996) Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative science quarterly* pp 116–145.

[65] Newell A, Simon HA, et al. (1972) *Human problem solving* (Prentice-Hall Englewood Cliffs, NJ) Vol. 104.

[66] Cyert RM, March JG, et al. (1963) A behavioral theory of the firm. *Englewood Cliffs, NJ* 2.

[67] Kahneman D, Slovic P, Tversky A (1982) Judgment under uncertainty: Heuristics and biases.

[68] Nelson RR, Winter SG (2009) *An evolutionary theory of economic change* (Harvard University Press).

[69] March JG (1994) *Primer on decision making: How decisions happen* (Simon and Schuster).

[70] Simon HA (1960) The new science of management decision.

[71] Simon HA (1957) *Models of man: social and rational; mathematical essays on rational human behavior in society setting* (Wiley).

[72] Burt RS (2005) *Brokerage and closure: An introduction to social capital* (Oxford university press).

[73] Latora V, Nicosia V, Panzarasa P (2013) Social cohesion, structural holes, and a tale of two measures. *Journal of Statistical Physics* 151:745–764.

[74] Coleman JS (1988) Social capital in the creation of human capital. *American journal of sociology* pp S95–S120.

[75] Granovetter MS (1973) The strength of weak ties. *American journal of sociology* pp 1360–1380.

[76] Granovetter M (2005) The impact of social structure on economic outcomes. *The Journal of economic perspectives* 19:33–50.

[77] Lin N (2002) *Social capital: A theory of social structure and action* (Cambridge university press) Vol. 19.

[78] Lin N, Cook KS, Burt RS (2001) *Social capital: Theory and research* (Transaction Publishers).

[79] Lingo EL, O'Mahony S (2010) Nexus work: Brokerage on creative projects. *Administrative Science Quarterly* 55:47–81.

[80] Burt RS (2009) *Structural holes: The social structure of competition* (Harvard university press).

[81] Burt RS (2000) The network structure of social capital. *Research in organizational behavior* 22:345–423.

[82] Burt RS, Guilarte M, Raider HJ, Yasuda Y (2002) Competition, contingency, and the external structure of markets. *Advances in Strategic Management* 19:167–218.

[83] Granovetter MS (1995) *Getting a job: A study of contacts and careers* (University of Chicago Press).

[84] Panzarasa P, Bonaventura M (2015) Emergence of long-range correlations and bursty activity patterns in online communication. *Physical Review E* 92:062821.

[85] Johnson S (2002) *Emergence: The connected lives of ants, brains, cities, and software* (Simon and Schuster).

[86] Estrada E (2012) *The structure of complex networks: theory and applications* (Oxford University Press).

[87] Barabási AL, Frangos J (2014) *Linked: the new science of networks science of networks* (Basic Books).

[88] Bar-Yam Y (1997) *Dynamics of complex systems* (Addison-Wesley Reading, MA) Vol. 213.

[89] Widrow B (1960) An adaptive adalineneuron using chemical memistors, 1553–1552.

[90] LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444.

[91] Rand DG, Arbesman S, Christakis NA (2011) Dynamic social networks promote cooperation in experiments with humans. *Proceedings of the National Academy of Sciences* 108:19193–19198.

[92] Mervis J (2015) Business decisions. *Science* 348:1190–1193.

[93] Kane TJ (2010) The importance of startups in job creation and job destruction. *Available at SSRN 1646934*.

[94] WhiteHouse (2010) Startup america. *Available at* `http://www.whitehouse.gov/economy/business/startup-america`.

[95] EuropeanCommission (2013) Entrepreneurship 2020 action plan. *Available at* `http://ec.europa.eu/enterprise/policies/sme/entrepreneurship-2020/`

index_ en. htm.

[96] Sorenson O, Singh J (2007) Science, social networks and spillovers. *Industry and Innovation* 14:219–238.

[97] Di Gregorio D, Shane S (2003) Why do some universities generate more start-ups than others? *Research policy* 32:209–227.

[98] Shane S, Stuart T (2002) Organizational endowments and the performance of university start-ups. *Management science* 48:154–170.

[99] Kalnins A, Lafontaine F (2013) Too far away? the effect of distance to headquarters on business establishment performance. *American Economic Journal: Microeconomics* 5:157–179.

[100] Saxenian A (1991) The origins and dynamics of production networks in silicon valley. *Research policy* 20:423–437.

[101] Ferrary M, Granovetter M (2009) The role of venture capital firms in silicon valley's complex innovation network. *Economy and Society* 38:326–359.

[102] Pena I (2002) Intellectual capital and business start-up success. *Journal of intellectual capital* 3:180–198.

[103] Ries E (2011) *The lean startup: How today's entrepreneurs use continuous innovation to create radically successful businesses* (Crown Books).

[104] Ledford H (2015) How to solve the world's biggest problems. *Nature* 525:308–311.

[105] Jacobs JA, Frickel S (2009) Interdisciplinarity: A critical assessment. *Annual review of Sociology* pp 43–65.

[106] Jacobs JA (2014) *In defense of disciplines: Interdisciplinarity and specialization in the research university* (University of chicago Press).

[107] Larivière V, Gingras Y (2010) On the relationship between interdisciplinarity and

scientific impact. *Journal of the American Society for Information Science and Technology* 61:126–131.

[108] Levitt JM, Thelwall M (2008) Is multidisciplinary research more highly cited? a macrolevel study. *Journal of the American Society for Information Science and Technology* 59:1973–1984.

[109] Chakraborty T, Tammana V, Ganguly N, Mukherjee A (2015) Understanding and modeling diverse scientific careers of researchers. *Journal of Informetrics* 9:69–78.

[110] Porter AL, Cohen AS, Roessner JD, Perreault M (2007) Measuring researcher interdisciplinarity. *Scientometrics* 72:117–147.

[111] Sayama H, Akaishi J (2012) Characterizing interdisciplinarity of researchers and research topics using web search engines. *PloS one* 7:e38747.

[112] Faust K (1997) Centrality in affiliation networks. *Social networks* 19:157–191.

[113] Freeman LC, Roeder D, Mulholland RR (1979) Centrality in social networks: Ii. experimental results. *Social networks* 2:119–141.

[114] Newman ME (2001) Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Physical review E* 64:016132.

[115] 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* 105:17268–17272.

[116] Radicchi F, Castellano C (2011) Rescaling citations of publications in physics. *Physical review E* 83:046116.

[117] Schummer J (2004) Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology. *Scientometrics* 59:425–465.

[118] Whitfield J (2008) Group theory. *Nature* 455:720–723.

[119] Shannon CE (2001) A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5:3–55.

[120] Brockmann D, Helbing D (2013) The hidden geometry of complex, network-driven contagion phenomena. *Science* 342:1337–1342.

[121] Newman ME (2002) Spread of epidemic disease on networks. *Physical review E* 66:016128.

[122] Colizza V, Barrat A, Barthélemy M, Vespignani A (2006) The role of the airline transportation network in the prediction and predictability of global epidemics. *Proceedings of the National Academy of Sciences of the United States of America* 103:2015–2020.

[123] Zhao L, Lai YC, Park K, Ye N (2005) Onset of traffic congestion in complex networks. *Physical Review E* 71:026125.

[124] Moreno Y, Pastor-Satorras R, Vespignani A, et al. (2003) Critical load and congestion instabilities in scale-free networks. *EPL (Europhysics Letters)* 62:292.

[125] Alessandretti L, Karsai M, Gauvin L (2015) User-based representation of time-resolved multimodal public transportation networks. *arXiv preprint arXiv:1509.08095*.

[126] Alessandretti L, Sapiezynski P, Lehmann S, Baronchelli A (2016) Multi-scale spatio-temporal analysis of human mobility. *arXiv preprint arXiv:1609.05514*.

[127] Yan G, Zhou T, Hu B, Fu ZQ, Wang BH (2006) Efficient routing on complex networks. *Physical Review E* 73:046108.

[128] Fronczak A, Fronczak P (2009) Biased random walks in complex networks: The role of local navigation rules. *Physical Review E* 80:016107.

[129] Latora V, Marchiori M (2007) A measure of centrality based on network efficiency. *New Journal of Physics* 9:188.

[130] Holme P, Saramäki J (2012) Temporal networks. *Physics reports* 519:97–125.

[131] Tang J, Musolesi M, Mascolo C, Latora V (2009) *Temporal distance metrics for social network analysis* (ACM), pp 31–36.

[132] Nicosia V, et al. (2012) Components in time-varying graphs. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 22:023101.

[133] Csardi G, Nepusz T (2006) The igraph software package for complex network research. *InterJournal, Complex Systems* 1695:1–9.

[134] Schult DA, Swart P (2008) *Exploring network structure, dynamics, and function using NetworkX* Vol. 2008, pp 11–16.

[135] Leskovec J, Sosič R (2014) Snap.py: SNAP for Python, a general purpose network analysis and graph mining tool in Python. (`http://snap.stanford.edu/snappy`).

[136] Peixoto TP (2014) The graph-tool python library. *figshare*.

[137] Kyrola A, Blelloch G, Guestrin C (2012) *GraphChi: large-scale graph computation on just a PC* (USENIX Association), pp 31–46.

[138] Szell M, Sinatra R, Petri G, Thurner S, Latora V (2012) Understanding mobility in a social petri dish. *Scientific reports* 2.

[139] Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: bringing order to the web.

[140] Latora V, Marchiori M (2001) Efficient behavior of small-world networks. *Physical review letters* 87:198701.

[141] Sabidussi G (1966) The centrality index of a graph. *Psychometrika* 31:581–603.

[142] Brandes U (2001) A faster algorithm for betweenness centrality*. *Journal of Mathematical Sociology* 25:163–177.

[143] Ramalingam G, Reps T (1996) On the computational complexity of dynamic graph

problems. *Theoretical Computer Science* 158:233–277.

[144] Lee MJ, Lee J, Park JY, Choi RH, Chung CW (2012) *Qube: a quick algorithm for updating betweenness centrality* (ACM), pp 351–360.

[145] Kas M, Carley KM, Carley LR (2013) *Incremental closeness centrality for dynamically changing social networks* (IEEE), pp 1250–1258.

[146] Nasre M, Pontecorvi M, Ramachandran V (2014) in *Mathematical Foundations of Computer Science 2014* (Springer), pp 577–588.

[147] Kourtellis N, De Francisci Morales G, Bonchi F (2014) Scalable online betweenness centrality in evolving graphs.

[148] Dijkstra EW (1959) A note on two problems in connexion with graphs. *Numerische mathematik* 1:269–271.

[149] Demetrescu C, Eppstein D, Galil Z, Italiano GF (2010) *Dynamic graph algorithms* (Chapman & Hall/CRC), pp 9–9.

[150] Hesse W, Immerman N (2003) Dynamic computational complexity. *University of Massachusetts Amherst*.

[151] Demetrescu C, Italiano GF (2001) *Fully dynamic all pairs shortest paths with real edge weights* (IEEE), pp 260–267.

[152] Frigioni D, Ioffreda M, Nanni U, Pasqualone G (1998) Experimental analysis of dynamic algorithms for the single source shortest paths problem. *Journal of Experimental Algorithmics (JEA)* 3:5.

[153] King V (1999) *Fully dynamic algorithms for maintaining all-pairs shortest paths and transitive closure in digraphs* (IEEE), pp 81–89.

[154] Demetrescu C, Italiano GF (2004) A new approach to dynamic all pairs shortest paths. *Journal of the ACM (JACM)* 51:968–992.

[155] Kas M, Carley KM, Carley LR (2014) An incremental algorithm for updating betweenness centrality and k-betweenness centrality and its performance on realistic dynamic social network data. *Social Network Analysis and Mining* 4:1–23.

[156] Demetrescu C, Italiano GF (2006) Experimental analysis of dynamic all pairs shortest path algorithms. *ACM Transactions on Algorithms (TALG)* 2:578–601.

[157] Bahmani B, Chowdhury A, Goel A (2010) Fast incremental and personalized pagerank. *Proceedings of the VLDB Endowment* 4:173–184.

[158] Zhang H, Lofgren P, Goel A (2016) Approximate personalized pagerank on dynamic graphs. *arXiv preprint arXiv:1603.07796*.

[159] Leskovec J, Lang KJ, Dasgupta A, Mahoney MW (2009) Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics* 6:29–123.

[160] Leskovec J, Huttenlocher D, Kleinberg J (2010) *Predicting positive and negative links in online social networks* (ACM), pp 641–650.

[161] Puzis R, Elovici Y, Zilberman P, Dolev S, Brandes U (2015) Topology manipulations for speeding betweenness centrality computation. *Journal of Complex Networks* 3:84–112.

[162] Michael TG, Tamassia R (2002) Algorithm design, foundations, analysis and internet examples.

[163] Gabow H (2003) Searching (ch 10.1). *Gross, JL; Yellen, J., Discrete Math. and its Applications: Handbook of Graph Theory* 25:953–984.

[164] Gambardella A (1994) The changing technology of technical change: General and abstract knowledge and the division of innovative labor. *Research Policy* 23:523–532.

[165] Walker G, Kogut B, Shan W (1997) Social capital, structural holes and the formation of an industry network. *Organization science* 8:109–125.

[166] Orsenigo L, Pammolli F, Riccaboni M (2001) Technological change and network dynamics: lessons from the pharmaceutical industry. *Research policy* 30:485–508.

[167] Leahey E, Beckman C, Stanko T (2015) Prominent but less productive: The impact of interdisciplinarity on scientists' research. *arXiv preprint arXiv:1510.06802*.

[168] Singh J, Fleming L (2010) Lone inventors as sources of breakthroughs: Myth or reality? *Management Science* 56:41–56.

[169] Bettencourt LM, Lobo J, Strumsky D (2007) Invention in the city: Increasing returns to patenting as a scaling function of metropolitan size. *Research Policy* 36:107–120.

[170] Klaus A, Yu S, Plenz D (2011) Statistical analyses support power law distributions found in neuronal avalanches. *PloS one* 6:e19779.

[171] Alstott J, Bullmore E, Plenz D (2014) powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one* 9:e85777.