Positive and negative connections and homophily in complex networks

Valerio Ciotti

A thesis submitted in partial fulfillment of the requirements of the Degree of the School of Mathematical Sciences
Doctor of Philosophy

2017

Declaration

I, Valerio Ciotti, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third partys copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

Signature: Valerio Ciotti Date: September 2017

Details of collaboration and publications: parts of this work have been completed in collaboration with Vito Latora, Pietro Panzarasa, Moreno Bonaventura, Vincenzo Nicosia, Andrea Capocci, Ginestra Bianconi, and Francesca Colaiori and are published in the following papers:

- V. Ciotti, G. Bianconi, A. Capocci, F. Colaiori, and P. Panzarasa (2015). Degree correlations in signed social networks. Physica A: Statistical Mechanics and its Applications 422: 25-39
- V. Ciotti, M. Bonaventura, V. Nicosia, P. Panzarasa, and V. Latora (2015). Homophily and missing links in citation networks. EPJ Data Science, 5(1), 1-14

Abstract

In this thesis I investigate the effects of positive and negative connections on social and organization networks, and the presence and role of homophily in networks of scientific collaborations and citations through the combination of methodologies borrowed from complexity science, statistics, and organizational sciences.

In the first part of the thesis, I study the differences between patterns of positive and negative connections among individuals in two online signed social networks. Findings suggest that the sign of links in a social network shapes differently the network's topology: there is a positive correlation between the degrees of two nodes, when they share a positive connection, and a negative correlation when they share a negative connection.

I then move my focus to the study of a dataset on start-ups from which I construct and analyse the competition and mobility networks among companies. Results show that the presence of competition has negative effects on the mobility of people among companies and on the success of the start-up ecosystem of a nation.

Competitive behaviours may also emerge in science. Therefore, in the second part of this thesis, I focus on a database of all papers and authors who have published in the American Physical Society (APS) journals. Through the analysis of the citation network of the APS, I propose a method that aims to statistically validate the presence (or absence) of a citation between any two articles. Results show that homophily is an important mechanism behind the citation between articles: the more two articles share similar bibliographies, i.e., deal with similar arguments, the more likely there is a citation between them.

In the last chapter, I investigate the presence of homophily in the APS data set, this time at the level of the collaboration network among scientists. Results show that homophily can be responsible in fostering collaboration, but above a given point the effect of similarity decreases the probability of a collaboration. Additionally, I propose a model that successfully reproduces the empirical findings.

Acknowledgements

Ever since I was little, when people asked me "what do you want to be?" I always responded, without any hesitation, "a scientist". I have always been fascinated in uncovering the hidden rules that govern nature. The decision to pursue a PhD was a straightforward one. I knew that to becoming a Doctor of Philosophy would have meant to learn how to properly "make" science, with all its positive and negative outcomes. But I did not know how complicated and hard it would be. Luckily, during these four years of study and scientific achievements, I have been surrounded by family, friends, and colleagues who made this path more enjoyable.

For my academic achievements, I would like to acknowledge my supervisors. In particular I would like to thank Vito for the support I received during my PhD. With his suggestions, bright ideas, and presence, I feel matured in my scientific approach to research. This helped to foster trust in myself and in my critical judgement towards problems faced on my scientific path. I also would like to thank Andrea Capocci, who first showed me the world of complex networks and who, with a lot of patience, taught me the basics of scientific research.

I want to thanks my colleague and friend Jacopo Iacovacci, who was there during the most complicated part of this path, proving that having a good friend by your side in time of need is priceless. I also want to thank the other people with which I spent a lot of time over the last four years in London, and whose encounters have contributed to make me the person I am today: Federico, Massimo, André, Trevor, Tom, and Isabella. Particular thanks go to Andrea Santoro and Iacopo Iacopini with whom I have spent the last months of my PhD and who were able to fill the gaps of those who left before me (which was a hard job!).

A special thanks goes to my friend Nils Haug, with whom I have also spent two years as flatmate, and who made me feel at home in a foreign country; because home is where somebody is waiting to eat dinner with you (I will never forget the spatzler and the butternut soup!).

Another special thanks goes to a person that I consider a colleague, a friend, and most importantly, a brother: Moreno Bonaventura. I have learned a lot from him, from both a scientific and human point of view. He opened the door towards a new way of looking at the world. He welcomed me into his start-up, giving me the opportunity to share with him incredible discoveries and results. I will be always grateful to him.

A final special thanks goes to Hannah-Maria, who represents the beginning of my revival, with whom I learned new and mature meaning of the word love. Her patience, maturity, and presence have undoubtedly contributed to make this path a sweet downhill.

Finally, I cannot forget the multitude of people that are always in my heart spread around the world: Pierpaolo, Alessandro, Nicola, Flavia, Alessia, and Enrico.

Last but not least, with deep gratitude, I wish to thank my family, without whom all of this could not even have taken place: Albertina, Giancarlo, Philippe, and Fabrizio.

Who I am today I owe in large part to them.

To my grandfather,
who I believe, is still watching over me, somewhere.

To my grandmother,
who watches over me, from Rome.

To my parents,
without which all this could not even take place.

And never forget.. *verba volant, scripta manent*

Contents

List of Figures

Chapter 1

Introduction

> "*Pleasures are dear and difficult to get.*
> *Feasting the eye, fat grapes hung in the arbour,*
> *That the fox could not reach, for all his labour,*
> *And leaving them declared, they're not ripe yet.*"
>
> — La Fontaine, The fox and the grape

> "*One can state, without exaggeration, that the observation of and the search for similarities and differences are the basis of all human knowledge.*"
>
> — Alfred Nobel

Understanding how individuals' behaviour is affected by social relations has always been one of the main concerns of social theories. In the literature, two main schools of thought can be identified that try to explain how economic action or individual behaviour depends on the underlying social structure. On the one hand, a prominent role has been played by the intellectual tradition that is firmly grounded in a microscopic or atomised point of view on the structure of society and human action. On the other, in contrast to this *undersocialised* view, an opposing perspective on human action has embraced a macroscopic or *oversocialised* approach to social behaviour [1].

The undersocialised point of view on atomised actors has been embraced primarily in classical and neo-classical economics [2–4]. This perspective is based on the utilitarian tradition according to which social structure and social relations are assumed *a priori* not to have any impact on the production or consumption of goods. On the contrary, social relations, bargaining, negotiation, and mutual adjustment have been seen as a burden that could undermine the smooth functioning of competitive markets.

The oversocialised conception of human action is based on the fact that each individual action is strongly affected by social norms and values, the broad cultural and social traditions, and the opinions and behaviour of other individuals [5–7]. Society itself creates the set of rules and social norms that are consensually accepted and followed by each individual. According to this perspective, conformity to rules is not perceived as an encumbrance, because it has already been internalised through processes of socialisation.

Even if *prima facie* both conceptions seem to be conflicting, as Granovetter [1] suggested,

> *"both have in common the conception of action and decision carried out by atomized actors." (p. 6)*

According to the microscopic point of view, actors aim to satisfy their own interests and are therefore unaffected by social relations; in the macroscopic point of view, norms and social rules have been created in common accordance, are adhered to and internalised by all actors, and are therefore only marginally affected by ongoing social relations.

If our aim is to understand how human beings act, we should abandon the implicit atomisation of actors. Social context inevitably has an impact upon individuals' decisions, but is in turn influenced by the way individuals behave, make decisions, and interact with one another. As also suggested by Granovetter [1], a way to fruitfully integrate the micro and macro perspectives can be found in the so-called *meso* point of view that regards individuals' social relations and behaviour as inherently embedded within a social structure. According to this integrated meso-perspective, individuals are enabled and constrained by the social structure within which they act, and, at the same time, they contribute towards its emergence by

interacting with one another. The mesoscopic point of view is what over the last few decades has contributed towards the emergence and development of the *social network perspective*. The studies carried out in this thesis will take this mesoscopic point of view.

1.1 Complex networks

Many systems can be represented as networks, in which nodes (or vertices) are connected through links (or edges). The concept of networks boasts a long intellectual tradition in the social sciences and beyond. Anthropologists, psychologists, sociologists, and molecular biologists have used network-based theories and methods since the 1950s. One of the first works in which the concept of network was proposed dates back to the 1930s [8].

Most networks have typically been included under the broad category of *complex networks* to emphasise the fact that their collective properties and behaviour at the global level are neither irreducible, predictable from, or explainable in terms of the properties of its constituent components (i.e., the nodes) [9–11]. It is the interaction among constituents that plays the role of the emergence mechanism that transforms patterns at the micro level into higher-level emergent patterns (e.g., the small-world properties or the presence of a community structure in social networks).

To gain an understanding of the macro properties of a networked system, the underpinning structure of links among the components thus needs to be fully taken into account. Moreover, complex networks spontaneously evolve over time without any centralised control, and typically display an internal order resulting from a self-organization prompted by local mechanisms at the node level. In most cases, this order may, in turn, underpin the emergence of a number of peculiar functional properties such as robustness against external attacks or internal errors. For example, even though cells are not designed by an external "architect", their metabolic network keeps on working and survives even during attacks or environmental changes [12].

The study of complex networks has been mainly the domain of a branch of discrete mathematics known as graph theory, and has attracted the interest of theoretical physicists as well as scientists from other disciplines only in recent times. The main aim of studying networks is to investigate the mechanisms governing their

topology and the dynamics of phenomena and processes taking place on it. Tools from statistical mechanics can offer a useful way for studying both network topology and dynamics [13]. A network-based approach to the study of a complex system is characterised by a special emphasis on the interactions among the elements of the system rather than on the detailed properties of the individual elements themselves. From this perspective, research on complex networks began with the aim of defining new concepts and measures to characterise network topology [13–15]. One of the main results has been the identification of a number of unifying principles and statistical properties that are shared by the majority of the real-world networks. This means that different systems can be viewed as different realisations of the same common principles. In network terms, for example, the spread of a computer virus can be compared to a process of flu spreading in a social environment or to information diffusion within an organization; similarly, hacking a router may generate the same effects of the extinction of a species in an ecosystem.

When the nodes of the network represent individuals, groups of people, firms, or organizations and the links between nodes represent the nature of the relationship between nodes, we are dealing with *social networks*. The social network perspective is distinctively characterised by the emphasis it places on the importance of social relationships among interacting actors and by the systematic attempt to express theories, models, and applications in terms of relational concepts [16]. In particular, there seems to be consensus among scholars that the social network perspective can be described in terms of the following paradigmatic orientations:

1. the analysis focuses on the relations between actors, instead of trying to sort actors into categories defined by their inner attributes;

2. social structure is operationalised in terms of relations among actors and is regarded as emerging from regularities or patterns generated by the interactions among actors;

3. behaviour is explained in terms of the structural context within which it is embedded rather than in terms of inner forces within actors; that is, patterned relationships among multiple actors are seen as jointly affecting (i.e., enabling or constraining) network members' behaviour;

4. structure is developed over different networks strata (i.e., multiplex or network of networks) that may or may not be partitioned into discrete groups: thus, it is not assumed *a priori* that tightly bounded groups are, intrinsically, the building blocks of the structure; and

5. analytic methods deal directly with the patterned, relational nature of social structure in order to supplement mainstream statistical methods that demand independent units of analysis.

In particular, a major conceptual building block upon which this intellectual tradition has developed a number of theories is the so-called *ego network*. This represents a network centred around an individual focal actor surrounded by neighbours. In an ego network, the focal actor is typically referred to as *ego*, while the set of actors with which ego is connected as *alters*. The ensemble of ego, its alters, and all ties between ego and alters as well as among alters is called the ego network.

1.1.1 The network perspective in organization theory

Recently, the term "network" has often been used in order to describe an organizational form [17–19]. At the same time, the concept of networking has become an attractive concept that underlines the importance for individual firms to create linkages with one another that can be used to their own advantage. According to Nohira [20], the recent use of the network perspective in the literature can be justified by three main reasons: (i) the emergence of the so-called *new competition* [21], which represents a new form of competition among small entrepreneurial firms that differs from the old form mainly because of lateral and horizontal exchanges and relations within and among firms; (ii) recent technological developments, which enable firms to better coordinate internal operations and inter-firm transactions; and (iii) the development of network analysis into a sound methodological apparatus, and its spread among scholars in academia, initially within the boundaries of the social sciences in the early 1970s [22] and subsequently across other interdisciplinary research domains.

The network perspective may help to identify the origin of power [17], to understand the factors that facilitate or impede the creation of a new venture, and to

examine strategies for the creation of inter-firm alliances [23, 24]. A brief yet comprehensive review of some of the major research streams in organizational network scholarship can be found in Borgatti [25]. A number of theories and concepts have been proposed and developed through the social network perspective, examples of which are *social capital* and *network organization* theory.

Social capital refers to the value that individuals, organizations, or firms can extract from the underlying social network of connections within which they are socially embedded [26]. One of the seminal works on social capital has been conducted by Burt [27], who identified social capital in the lack of ties among an actor's alters. Burt typically refers to this topological configuration as *structural holes*. He argues that the spanning of structural holes can be regarded as the social mechanism that underpins Granovetter's theoretical argument on the strength of weak ties [28].

The other network-based concept that has witnessed increasing popularity over recent years is the so-called "network organization". This includes organizational forms characterised by repetitive exchanges among organizations based on trust and embedded social relationships aiming to protect transactions and reduce costs [29–31]. Scholars suggest that as commerce become much global and highly competitive, both markets and hierarchies show their limitation by displaying inefficiencies [19, 32]. A network organizational form has emerged to face these limitations, able to balance the flexibility of markets with the predictability of traditional hierarchies.

In this thesis I will focus my attention to a particular organization network: the network of start-ups. In particular, I will evaluate the effects of competition on both the mobility of employees between companies and on the success of the nation in which start-ups are located. I will create two networks of start-ups: one in which two start-ups are connected if there is a relationship of competition among them, and a second in which two companies are connected if there has been, over time, an exchange of employees between them. For both analysis I will make use of a network approach.

1.2 Signed social networks

One of the main building blocks of my research is the concept of *signed social networks*. A social network is "signed" when relationships can have either a positive or

a negative connotation. For instance, positive relationships may refer to friendship, love, trust, collaboration, or advice. By contrast, examples of negative relationships are enmity, hatred, distrust, or competition. In such cases, it is possible to associate each link between nodes with either a positive symbol, such as "+", or a negative one, such as "−". In principle, one could also consider varying degrees and combinations of the positive and the negative relationships, but for the sake of simplicity in this thesis I shall consider as signed social networks only those networks where each link has been associated with a positive or a negative sign. The absence of a link between any two actors means that the actors nurture neither a positive nor a negative feeling towards each other.

One of the seminal theoretical achievements that marked the study of signed social networks over the last decades is *balance theory*. Originally proposed by Heider [33], balance theory is based on the notion of balanced relations: a relation between two actors is defined as balanced if the signs of the two links connecting the actors in both directions are the same, i.e., the two actors are connected through a bidirectional and reciprocal link associated with the same positive or negative connotation in both directions. Empirical work has consistently reported that an abundance of social relations are indeed balanced. As a result, over the years scholars have proposed a number of theories that attempt to explain this recurrent empirical regularity, and in particular why and how social relationships tend to develop into a balanced state.

Balance theory is based on the psychological concept of *cognitive dissonance* proposed by Festinger [34]. The idea is that people have a motivational drive to reduce the number of dissonant elements (e.g., incoherent ideas or behaviour) by altering their cognitive world, modifying the environment or replacing old beliefs with new consistent ones. Individuals are engaged in a process Festinger defined as *dissonance reduction* [34]. Balance theory suggests that if people recognise a set of cognitive elements as being a system, they will have a preference to maintain a balanced state among these elements, and therefore will tend to reduce the cognitive dissonance afflicting them. One of the most famous example of cognitive dissonance is given in the fable "*The Fox and the Grapes*" by Aesop. In the story a fox sees a bunch of grapes inaccessible for its height; at the end of the story, the fox convinces itself that the grapes are probably not worth eating. This is a clear example of

the fact that when one desires something and finds it unattainable, it is possible to reduce the generated dissonance by re-evaluating it and changing one's own mind.

Heider's [35] theory was initially developed in order to understand individual's cognition or perception of social situations. Heider focused on a single individual, and in particular he was concerned about the individual's attitudes or opinions and their relations with the attitudes or opinions of other individuals. Specifically, Heider [35] claimed that:

> *"In the case of two entities, a balanced state exists if the [ties] between them [are] positive (or negative) in all aspects [...] In case of three entities, a balanced state exists if all ties are positive in all respects, or if two are negative and one positive." (p. 110)*

For example, we can consider two individuals, and their opinions about a statement. If both actors are connected by a positive relation, then they should react similarly to a given statement (i.e., both of them should either oppose or favour it). We have a balanced state if the two actors act in the same way, and perceive this to be the case. By contrast, if they hold conflicting attitudes towards the same statement, we have an unbalanced state, and each of the two individuals involved will perceive a cognitive dissonance.

The concepts proposed by Heider have been operationalised and mathematically represented using graph theory by Cartwright and Harary [36] and Davis [37]. In particular, cognitive balance has inspired the development of *structural balance theory*, which does not focus on the individual, but rather on a set of individuals or groups. According to the structural balance theorem, once all relationships within a network are balanced, all network members either become friendly with one another, or divide themselves into two opposing camps [36]. Subsequently Davis [37] provided a generalisation of the structural balance theorem to cases in which individuals split into more than two mutually hostile groups.

The central idea behind structural balance is that configurations of signed local groups of actors in a network containing positive and/or negative ties are socially and psychologically more stable than other non-balanced configurations, and are therefore more likely to be found in real-world social networks. To illustrate this point, the four triads at the top of Fig.1.1 are balanced and are allowed in a struc-

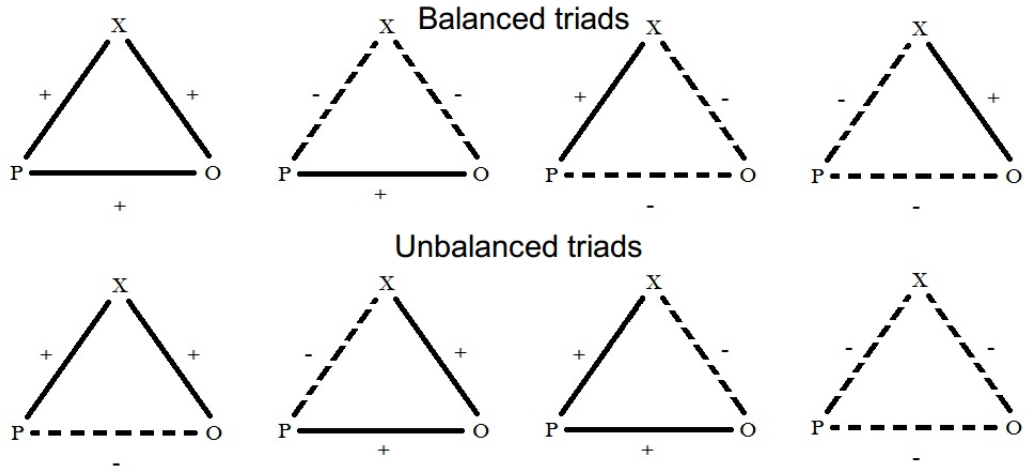

Figure 1.1: **Examples of balanced and unbalanced triads**

turally balanced network, while the four triads at the bottom are unbalanced and unlikely to occur in reality. For instance, real-world social networks have been found to be characterised by an abundance of positive triads in which triplets of actors contain either three positive links, or one positive and two negative links. This has been summarised with the two-fold hypothesis that "the friend of my friend tends to be my friend", and "the enemy of my enemy tends to be my friend". Conversely, social networks have been found to exhibit paucity of unbalanced connected triads, namely triads containing an odd number of negative links. In turn, this has been referred to as the hypothesis that the enemy of my enemy is likely to be my enemy, and the friend of my friend is likely to be my enemy.

Structural balance has been tested in many empirical applications, including the study of international relations among nations, where relations typically refer to political alliances during times of warfare [38, 39], and the study of politicians or community elites as actors involved in positive and negative relations [40]. The goal of these studies has been to examine the underlying social structure, and to uncover how much tension was caused by the interplay of negative and positive relationship among subsets of actors.

In this thesis I will explore in which measure the presence of balance (or unbalance) affects the topological structure of signed social networks. In particular, I will

propose a model in which I will make use of structural balance in order to reproduce empirical findings concerning the topological differences generated by positive and negative connections among social actors.

1.3 Homophily and heterophily: Revisiting their interplay

Balance theory has consistently been used to derive a set of testable hypotheses regarding how individuals create, modify, or sever social relationships in order to avoid or mitigate dissonant feelings. As such, balance theory has contributed towards a better understanding of how social networks evolve over time as a result of the way individuals interact with one another.

Over the last few decades, the literature has suggested a number of network growth mechanisms to explain how social relationships are forged and severed over time. For example, the principle of *social embeddedness*, proposed by Coleman [41], refers to the hypothesis that, because social capital originates from closed social structures, individuals tend to cluster into tightly knit groups that are rich in third-party social relationships. In a similar vein, the mechanism of *triadic closure* [37] formalises the idea that, given a triplet of connected nodes, such that node i is connected to node j and node l, it is more likely that nodes j and l are also connected with each other than would be the case if nodes j and l did not share i in common. Another well documented network growth mechanism is *cumulative advantage*, namely the hypothesis that once an individual gains a small advantage over other individuals, that advantage will compound over time into an increasingly larger advantage. This effect has been widely investigated in many empirical domains, and is typically referred to by the literature as the principle that "the rich get richer" [42] or, more recently, as the principle of "preferential attachment" [14].

Among these network growth mechanisms, a key role is played by homophily, the principle that similarity breeds connection ("birds of a feather flock together") [18, 43]. In particular, homophily can be seen as the general network principle that subsumes a number of more specific empirical regularities, such as the tendency of social networks to exhibit degree correlations. In fact, if the creation of a relation-

ship is based on the principle that similar individuals are more likely to connect with each other than dissimilar ones, then the resulting network will partition into a number of heterogeneous communities, each composed of similar individuals that share similar ideas, beliefs or interests. Individuals are thus expected to forge most of their social relationships with other individuals within their own community, while only a minority of links will be established with other individuals that belong to different communities. Because the degree (i.e., number of connections) of nodes is constrained by the size of the community to which they belong, the implication of homophily and the subsequent partition of the network into communities is that the average number of connections (i.e., the degree) of an individual's nearest neighbours is likely to be correlated with the individual's degree. This leads precisely to assortativity, i.e., the network property concerning the positive correlation between a node's degree and its neighbors' degree. We can conclude that the network community structure is a key factor in determining assortativity. Being homophily one of the main determinants of the emergence of communities [44], in order to properly understand how and why assortativity occurs in social networks, it is fundamental to investigate the presence of homophily in social networks.

Homophily has been empirically documented in a variety of domains, including marriage, friendship, work advice, social support, and information transfer. McPherson *et al.* [43] distinguishes two types of homophily: (i) *status homophily*, which refers to the hypothesis that social relationships are likely to occur between individuals that are similar in terms of a number of socio-demographic characteristics, such as race, ethnicity, gender, age, religion, formal or informal status, and (ii) *value homophily*, which is expressed in terms of similarity among individuals based on cognitive attitudes, including values, beliefs, goals, and the moral and ethical principles that are supposed to affect or guide behaviour.

Geographic propinquity, affiliations with families and organizations, and isomorphic positions in social systems have been found to create suitable contexts in which homophilous relationships are likely to emerge. Among the various forms of value homophily, cognitive homophily suggests that social interactions between individuals arise precisely as a result of their similarity and convergence on the same ideas, beliefs, interests, and mental attitudes. Moreover, ties between non-similar individuals are likely to dissolve at a higher rate than ties between similar ones, which

sets the stage for the formation of cliques (i.e., localised positions) within the social space [43].

Homophily in organizations reflects the fact that usually individuals are likely to create and belong to many small groups where a given feature is shared among the members of the same group. For example, a person may be a football supporter, a student, and of Italian nationality, and may find himself drawn towards people of each of these different groups. Such behaviour in organizations may lead to problems of social fragmentation. Employees that share similar characteristics may spend most of their time with each other, disregarding the need to cultivate relationships with the rest of their (dissimilar) colleagues [45]. In this sense, homophily is a double-edged sword: it induces the creation of tightly knit communities and reinforces relationships and trust among people within the communities, but it also leads to a scarce circulation of new information and knowledge beyond and across the boundaries of the communities [46].

Even if homophily has placed emphasis on similarity and its consequences for tie creation, there has been also an substantial body of literature in the social sciences that has underlined the importance of forging relationships between dissimilar people. The principle that describes the tendency of dissimilar individuals to interact with one another is typically referred to as *heterophily*. One of the leading exponents of the heterophily theory has been Georg Simmel [47], increasingly regarded as one of the pioneers of network analysis. The benefits that individuals can extract from their dissimilarity have fruitfully been articulated by Simmel in his sociological theory of the *stranger*. Simmel describes the stranger as someone who "*comes today and stays tomorrow*" [47, p. 126]. That is, the stranger does not have or know his or her role in society; he or she is near, yet far away from the group to which he or she belongs. People can trust the stranger because he or she knows no one else in that specific society, and so they can feel free to talk to her in confidence and trust that he or she will not judge them or reveal their personal accounts to anyone else. The role of the stranger is to bring innovation, news, and information to the groups to which he or she is connected. He or she brokers relations between the groups within which people dwell and the groups with which he or she maintains distant relations. The strength of the stranger is in his or her weak ties or, in other words, in the inter-group relationships through which he or she can channel and spread

information. From this perspective, and drawing on this theoretical argument, it has been conjectured that an individual may be more likely to connect with others that are dissimilar in order to gather new information or resources than with those that are similar, who cannot offer information or resources that the individual does not already possess.

The two main works based on heterophily are Granovetter's contribution of the strength of weak ties [28,46] and Burt's theory of structural holes [27,48,49]. Those two contributions are deeply connected to each other. They are both concerned with theorising the antecedents of social capital, which is premised on the idea that investments in social relations yield expected returns in the marketplace, including the community, the economic, financial, political, and labour markets. In a social network, the intensity of a tie among two individuals may assume different values that are proportional, as Granovetter has suggested, to "the amount of time, the emotional intensity, the intimacy, and the reciprocal services which characterize the tie" [28, p. 3]. Strong ties are conducive towards local connections and sustain the creation of closed, trustful but separate groups of individuals [50]. Information within these groups tends to be rather identical and repetitive, thus losing its value for the members of the groups. The only way in which these members can obtain new information is by creating new bridging ties that enable them to reach other individuals that belong to different groups. These bridging ties are *emotionally* weak, in that they tend to be characterised by low intensity, frequency and intimacy. However, they are *structurally* strong, in that they enable actors to extract social capital from their underlying network, and on a global scale they allow information to flow throughout the network.

Building upon Granovetter's work, Burt has investigated how social capital can originate from brokerage opportunities associated with structural gaps in the network. Burt has defined a structural hole as the "separation between non-redundant contacts, a relationship of non-redundancy between two contacts, a buffer" that enables the two contacts to "provide network benefits that are in some degree additive rather than overlapping" [27, p. 18]. A broker can exploit his or her structural position as the gatekeeper between contacts at the opposite sides of the hole. There are two types of benefits associated with this position: (i) information benefits that originate from the fact that, in structures rich in structural holes (i.e., open struc-

tures), connections with otherwise disconnected individuals or groups tend to be weak [28], and are likely to bring the focal actor (i.e., the broker) closer towards people with different ideas, interests, new opportunites and perspectives; and (ii) social control benefits that result from the broker's ability to negotiate his or her relationships with otherwise disconnected others and turning their "forces combined against him into action against one another" [47, p. 162].

There is a clear analogy between Burt's conception of the broker and the Simmelian "stranger", and both are conceptually consistent with Granovetter's idea on the benefits of weak ties for social capital. Both the broker and the stranger are actors that are weakly connected to different, otherwise disconnected groups of people, and for this reason may offer and receive new information, as well as fresh insights, and new perspectives. In this sense, both concepts of the broker and the stranger point to the salience of heterophily as a principle governing social relationships. If individuals want to extract social capital from their underlying social network by gathering and combining different pools of new information, it is more likely that they will forge relationships with dissimilar others than similar ones.

1.3.1 Social dependence

I have shown two apparently opposing social mechanisms responsible for the creation of social ties: one based on the hypothesis that the more similar we are the more likely we are to connect with one another; the other based instead on the hypothesis that we are inclined to connect with dissimilar others. In what follows, my aim is to draw on the relevant literature in order to explain the dynamic interplay between the two principles of homophily and heterophily. I shall do so by regarding heterophily as a form of social dependence between individuals. In so doing, I shall put forward the hypothesis that homophily and heterophily go hand-in-hand, such that if individuals are similar, they are more likely to interact than if they are dissimilar (i.e., homophily), but at the same time individuals seek connections precisely in order to satisfy each others' needs through the dissimilar resources they possess and can offer (i.e., heterophily). In this sense, while homophily remains expressed in terms of socio-demographic and cognitive characteristics, heterophily is predicated in terms of the material (e.g. goods) or immaterial (e.g. information) resources that

individuals possess and exchange.

This idea is not new. organizational ecologists have long suggested that similarity can also lead to competition for scarce resources. According to this strand of research, similarity and competition go hand-in-hand: high concentration of similar organizations can lead to competition for scarce resources [51–54]. For example, Ahuja *et al.* [51] have investigated the alliances involving 97 global chemical firms. They have argued that poorly embedded firms are more likely to participate in ties characterised by social asymmetry than in ties characterised by structural homophily. Indeed, beyond a certain threshold, a firm's centrality (i.e., the number of connections to other firms) was found to create a disincentive for potential collaborations, thus diminishing the likelihood that two firms with an equally large value of centrality will form an alliance.

In the organizational literature, it is possible to identify two chief forms of social interdependence [55]: one that originates from collective action and the *horizontal* sharing of common resources [56]; the other that can be explained in terms of *vertical* forms of exchanges or transactions among units [54, 57].

Among the horizontal forms of social dependence, scholars have identified the so-called *pooled interdependence*. This refers to those forms of dependence arising from the fact that all individuals involved provide a contribution to a common achievement, or make use of common resources, and in turn can benefit on a proportional basis. Considering as an example the different departments of the same organization; in this type of interdependence, each organizational department carry out separate tasks. Even though each department work independently and do not directly depend on each other, in the "pooled interdependence model", each does contribute individual pieces to the same overall puzzle. This kind of social interdependence is usually important for firms that share the same market target and whose activities rely upon the same pooled resources, such as technologies, competencies or administrative structures.

Another type of horizontal interdependence is the so-called *intensive interdependence*. This form has been proposed by Thompson [54] in his seminal work on social dependence. This kind of interdependence is based on the joint cooperation of specialised actors that need to share complex knowledge in order to solve a common problem. A suitable example is the medical team during surgical interventions.

The other major type of social dependence is the vertical one that regulates exchanges among actors. The primary example of this form is referred to by the literature as *sequential interdependence*. This represents a form of connection between two activities such that the output of one activity is the input for the other. It represents the simplest case of transactional interdependence, where there is a transfer of goods or services through a given interface from one activity to another [57].

Social dependence has been extensively investigated in organizational theory, and has been the subject of long-standing debates among scholars, especially those concerned with power in organization. For instance, Emerson [58], one of the preeminent scholars interested in power, has suggested that social relations commonly entail ties of mutual dependence between parties. This means that actor i depends upon actor j if j has the appropriate resources or can help i to achieve his or her own goals. Usually the dependence is found to be mutual, i.e., i must be in a position to offer j something useful to j's satisfaction. The power of i over j is typically defined as the extent to which j is dependent on i [59]. Thus, actors who are able to control desired resources increase others' dependence on them and, through the process of exchange, are able to acquire the necessary resources and bring about the outcomes they desire [60, 61].

In this thesis I draw on the concept of social dependence. The aim is to develop a network model able to replicate the dynamics of tie creation that combines and extends the homophily and heterophily/ecological arguments through the concept of interdependence. This will enable me to investigate the non-linear effects of increasing degrees of similarity on the probability of tie creation. My hypothesis of the non-linear effects of homophily on tie creation can be described as follows; it is reasonable to expect that the probability and strength of an interaction increases with similarity between the interacting nodes, as suggested by the principle of homophily. However, this increase is expected to occur only up to a certain critical threshold value of similarity. Above this threshold, the effects of similarity reverse: individuals highly similar are less likely to provide one another with the information and resources they are looking for, and will thus direct their attention to other less similar partners. In this sense, the model I will propose will be rooted in the interplay between homophily and social dependence, and will formalise the hypothesis that socially interdependent individuals that are too similar to each other are unable to

satisfy each other's objectives, and thus tend to avoid their interaction. In so doing, my aim is to develop a theoretical argument and a corresponding analytical model that dovetails the interplay between the principles of homophily and heterophily. Ultimately, my work will also contribute towards bridging the theoretical divide between these two apparently opposing principles that the literature has proposed to explain the way social relationships are created and develop over time.

1.4 Outline of the thesis

In this thesis I will investigate the effects of positive and negative connections on social and organization networks, and the presence and role of homophily in networks of scientific collaborations and citations. I will start by analysing to which extent connections with a positive or a negative nature shape the network topology of two online social networks and I will propose a model based on balance theory intended to reproduce the empirical findings.

Positive and negative connections may also be found in other type of networks, such as organization networks, in which nodes represent companies or groups of people. In particular, I will study the competition among start-ups. I will create the competition network among start-ups, where nodes are start-ups and a connection exists if there is a relationship of competition among two companies. Making use of network techniques, I will quantify the effects of competition on the mobility of employees among start-ups and on the success of national ecosystems of start-ups.

Competitive behaviours may appear also in the scientific domain. One way to detect competition can be done by looking at the absence of a relevant citation among two scientific papers. In fact, citations in science are an important instrument to affirm the appreciation of a scholar's work. Through the analysis of the citation network among scientific papers published in the American Physical Society (APS) journals, I will propose a method that aims to statistically validate the presence (and the absence) of relevant citations. Specifically, I will show that citations follow homophily: the more similar the bibliography of two papers is, the more likely we will find a citation between them.

Finally, I will conclude this thesis with a study on the collaboration among scientists in the APS dataset. I will define a measure to quantify a scientist's scientific

interests and I will analyse the evolution of his or her interests over time. Based on this result, I will propose a measure of scientific similarity in order to test the interplay between the two opposing mechanisms of homophily and heterophily in forging scientific collaboration among scientists.

1.4.1 Thesis structure

The following chapters of the thesis are organized as follows:

Chapter 2 presents my study on the differences between positive and negative connections in social networks. In particular, I will analyse two online social networks (epinions.com and slashdot.org) in which links between individuals can be either positive (trust or friendship) or negative (distrust or enmity). I will detect differences between positive and negative connections by looking at the degree correlations of each node in respect to the degree of their neighbours. Findings indicate that, when the sign of links is ignored, both networks are assortative, i.e., each of the two networks nodes with similar degree tend to connect with each other. This result is in agreement with previous analyses that show social networks as characterized by an assortative trend as compared to other type of networks (e.g., technological and biological ones), which on the contrary present a disassortative trend, i.e., the tendency of nodes with dissimilar degree to connect with each other. To study the impact of the sign of links on degree correlations, from both networks I will extract the positive and negative subnetworks composed only by links of the same sign. Results indicate that the sign of links has some bearing on the degree correlations observed in social networks: the positive subnetworks are assortative while the negative ones are disassortative. To shed light on this result, I will then propose a network model in which I assign each node to one of two mutually exclusive groups, and associate a positive sign to connections between nodes of the same group and a negative sign to connections between nodes of different groups. Results of numerical simulations are in accordance with the empirical findings when a combination of three different conditions is met. I will investigate the role of each of these conditions and different combinations of them, and extend the analysis by studying the case in which the global unsigned network is disassortative and nodes can be allocated to three or more mutually exclusive groups.

Chapter 3 is devoted to the investigation of the effect of competition between start-ups on the mobility of people between companies, and on the success of the set of start-ups located in each nation. In the first two sections I will provide a detailed description of the dataset and the methodology used to construct the network of markets (or industry sectors) in which each company is involved. Unfortunately, markets that are assigned to each company do not belong to any hierarchical category, which makes it difficult to use at that level of detail. I will propose a methodology based on a combination of network techniques to define *macro market categories* to which each market will be assigned. I will then propose a measure to asses the differences and similarities between national ecosystems start-ups based on the activities of start-ups in each macro market category. In the third section I will construct two start-up networks, namely the network of declared competitors among start-ups and the mobility network of employees among start-ups. By studying the overlap between these two networks it is possible to asses the effects of competition on mobility. Results show that the presence of competition negatively impact both the mobility of people between companies and the success of the national ecosystem to which the start-ups belong.

Chapter 4 casts light on the salience of homophily, namely the principle that similarity breeds connection, for knowledge transfer between papers. To this end, I will asses the degree to which citations tend to occur among papers that are concerned with seemingly related topics or research problems. Through analysis of the citation network among physicists that have published in the American Physical Society (APS), I will propose a method that suggests the presence of relevant citations. In the first two sections I will give an overview on the studies related to citation networks and I will describe the APS dataset. In the third section I will present the methodology used to quantify and assess the statistical significance of the similarity between any two articles published in APS. Based on this measure, I will evaluate the absence of relevant citations or the presence of irrelevant ones. Results show that the more two articles share similar bibliographies, i.e., treat similar arguments, the more likely there is a citation between them. By assuming the presence of relevant missing citations as a lack of knowledge flow, I will propose to rank both different areas of physics and the APS' journals based on the lack of knowledge flow between papers.

Chapter 5 is devoted to the study of the evolution of physicists' careers over time and on the analysis of social mechanisms in forging collaboration among them. In the first section I will provide a detailed description of different tie creation mechanisms. In the second section I will present the dataset and the measure used to evaluate the evolution of a physicist's interests and specializations throughout his or her career. The third section will focus on the study of the collaboration network over time among physicists. In particular, I will look at whether the presence of a collaboration is driven by scientific similarity among physicists. Results show that the more two scientists are scientifically similar, i.e., specialised in similar topics, the more likely they are to collaborate. This is true up to a given threshold, above which the probability of a collaboration decreases. In section four, I will put forward the hypothesis that this non-linear effect is driven by the presence of two opposing forces that simultaneously act on two different levels: homophily and social dependence. The combination of these two effects creates the reversed "U-shaped" trend that emerges in the results. I will justify the hypothesis of the combination of these two effects through the use of a model able to reproduce the empirical findings.

Finally, **Chapter 6** is devoted to conclusions and discussion around possible future work.

Chapter 2

Signed social networks

"What loneliness is more lonely than distrust?"

— George Eliot

"Friends come and go but enemies accumulate."

— Arthur Bloch

Balance theory has been studied both locally, with an emphasis on dyadic relationships [35], and globally, with a prominence on the whole network and its partition into distinct groups [36, 37]. The work presented in this chapter is concerned with the global implications of balance. In particular my aim is to investigate the degree to which the presence of balance (or unbalance) affects relevant topological properties of social networks. Among these properties, one way to quantify the effect of balance is by evaluating the network degree correlations.

Degree correlation is a network property that captures the extent to which a node is likely to connect with other nodes that have a similar number of connections (i.e., degree). In a social network, if an individual is connected with others that have the same number of friends, the nodes' degrees are positively correlated with the degrees of their neighbours and the network is said to be assortative. In this case, individuals with many friends are connected with individuals with many friends, and individuals with few friends are connected with individuals with few friends. If instead individuals with many friends are connected primarily with individuals with few friends and vice versa, the nodes' degrees are anti-correlated with their neighbours' degrees and the network is said to be disassortative.

Focusing on degree correlations is a non-trivial task and has important implications for a better understanding of social networks. Indeed it has been showed by scholars that what makes social networks different from other networks types are two distinctive empirical regularities [62, 63]. First, in social networks actors that share a common friend are likely to connect to each other and form closed triadic relationships (which in turn contributes to an high clustering coefficient); second, social networks tend to be assortative, i.e., they have been found to exhibit positive degree correlations.

In non-social networks, nodes with a common neighbour are less likely to connect to each other than in social networks (which contributes to a low clustering coefficient), and the degrees of connected nodes tend to be anti- correlated (i.e., the networks are disassortative) [62]. Thus, by uncovering the implications that structural balance has on degree correlations, I shall make a step forward towards a better understanding of the antecedents of one of the structural properties that have long been regarded as distinctively characterising social networks.

A thorough analysis of the literature and the empirical work so far conducted on degree correlations suggests that correlations in negative social networks, i.e. networks in which the relationship between actors is mediated by negative connotations, still remain to be investigated. Negative social networks may indeed exhibit correlation patterns that differ from those detected in positive social networks, and, as a result, a relation may exist between the sign of the links and the type of degree correlations of a social network. Do individuals who distrust many others tend to distrust each other, or do they channel their negative feelings towards other individuals who distrust only very few others? Whether negative social ties, such as distrust or hatred, tend to be forged primarily between actors characterised by similar or dissimilar connections still remains largely unexplored. This chapter is devoted to rectify this shortcoming. In particular, once I have uncovered the extent to which the degree correlations of negative social networks differ from the correlations of positive ones, I shall study how this divergence is likely to be due to the presence or absence of structural balance in networks where positive relationships are intermingled with negative ones.

The outline of the chapter is as follows. In Section 2.2, I introduce two signed online social networks, and examine the degree distributions and correlations of

the positive and negative sub-networks extracted from the data. In Section 2.4, I propose a generative model of signed networks that polarize into two mutually exclusive groups of nodes. Section 2.4.1 focuses on the case of random networks with binomial degree distributions, whereas Section 2.4.2 deals with more realistic cases of networks with power-law degree distributions. Finally, in Section 2.5 I extend the modelling framework to networks in which nodes can split into three (or more) hostile groups. In Section 2.6, I summarize my findings and discuss their implications for research on signed complex networks.

2.1 Introduction

Over the last few years, an increasing interest in the study of social networks has prompted physicists, mathematicians and computer scientists to join sociologists in their endeavours to develop network models concerned with the antecedents, structure, and evolution of social interaction [14, 64, 65]. Recent studies have indicated that social networks across many empirical domains display the typical signature of complex networks, namely the long-tailed distribution of the degrees of nodes [14]. In addition to this, an attempt has been made to uncover the distinctive structural features and empirical regularities that distinguish social networks from other types of complex networks. While in most real non-social networks degrees of neighboring nodes tend to be anticorrelated, research has suggested that social networks tend to be characterized by the opposite correlation pattern [66, 67]. The tendency of nodes with similar degree to connect with each other is often referred to as "assortative mixing by degree", and has been observed in a number of social networks, including very large-scale online social networks such as Facebook and Twitter [68].

A variety of models have been proposed by physicists, sociologists and computer scientists to explain these distinctive properties of social networks. For instance, assortative mixing has been related to the underlying community structure of social networks [67]. More recently, assortative mixing has been explained in terms of transitivity [69] and homophily [70]. Sociologists have also uncovered distinctive interaction patterns within social signed networks in which relationships can have a positive (e.g., trust and friendship) or negative (e.g., distrust and enmity) connotation. In particular, the theory of "structural balance" has long suggested that, in

undirected signed social networks, individuals embedded within closed triads tend to minimize cognitive tension: an individual tends to befriend a friend's friend, distrust a friend's enemy, befriend an enemy's enemy, and distrust an enemy's friend [36,71].

In this chapter I focus my attention on the emergence of degree correlations in signed networks, and how these correlations can be used to predict the sign of links in cases where it is not known or cannot be assessed directly. Indeed, despite the ubiquity and salience of negative relationships in a wide range of social systems, the detection of mixing patterns by degree has been confined primarily within the domain of unsigned networks or simply networks in which nodes were assumed to be connected through positive links (e.g., scientific collaboration networks and interlocking directorate networks [67]). However, negative networks may exhibit correlation patterns that differ from those detected in positive networks. Do individuals who distrust many others tend to distrust each other, or do they channel their negative feelings toward other individuals who distrust only very few others? To address this problem, here I propose a class of simple models that help uncover the relation between the sign of links and the type of degree correlations characterizing a network.

2.2 The data

I analyze two online social networks. The first is the network formed by the users of Epinions (*www.epinions.com*), a website for user-generated reviews of various products. Registered users of Epinions can declare their trust or distrust toward one another, based on the comments they post. The second social network is formed by the users of Slashdot (*www.slashdot.org*), a website devoted to the discussion of technology-related news, and in which the Slashdot Zoo feature enables users to tag one another as "friends" or "foes". In both Epinions and Slashdot, connections are directed and signed. The meaning of the sign of links is similar: a positive link means that a user endorses another user's comments, whereas a negative one means that a user dislikes another user's comments. Both network datasets are available from the Stanford Network Analysis Project website [72].

Table 2.1 reports the number of nodes and links in the datasets [73,74]. Epinions is composed of $131,828$ nodes and $841,372$ directed links. In particular, $717,667$ of these links (i.e., 85.00%) are positive and represent the trust users accord to each other. Moreover, links connecting $130,162$ (i.e. 15.47% of all links) pairs of nodes in Epinions are reciprocated, of which only 1.8% are characterized by a combination of a positive and a negative signs (i.e., node i points positively to node j, and j points negatively to i). The Slashdot social network is composed of $82,144$ nodes and $549,202$ links, $425,072$ (i.e. 8.87% of all links) of which are positive (i.e., 77.40% of the total number of links). Moreover, $48,721$ pairs of nodes are connected through reciprocated links, of which only 4.0% are characterized by different signs.

To study the impact of the sign of social relationships on the network topology, for each network dataset I filtered out and isolated the positive and the negative subnetworks composed only by reciprocated links of same sign (see Fig.2.1). The choice of considering only reciprocated links helps to identify in a clear way the nature of the relationship among two individuals. Moreover, it allows to consider the whole network as undirected and to study the implications that structural balance (which is defined on undirected networks) has on degree correlations of the overall signed network. In particular, from the Epinions social network two signed subnetworks were extracted: the "trust" and "distrust" subnetworks in which all links are positive and negative, respectively. Similarly, I created the Slashdot "friend" and

	Epinions	Slashdot
Nodes	131,828	82,144
Links	841,372	549,202
Positive	717,667	425,072
	(85.30%)	(77.40%)
Negative	123,705	124,130
	(14.70%)	(22.60%)
Reciprocated	130,162	48,721
	(15.47%)	(8.87%)

Table 2.1: Nodes and links in Epinions and Slashdot.

"foe" subnetworks.

These four signed subnetworks are characterized by a power-law distribution $p(k) \simeq k^{-\alpha}$, with an estimated value of the coefficient α of approximately 2.35 (see Fig. 2.2). This value is similar to the one estimated for the power-law distribution of the unsigned network with reciprocated links (see inset of Fig. 2.2).

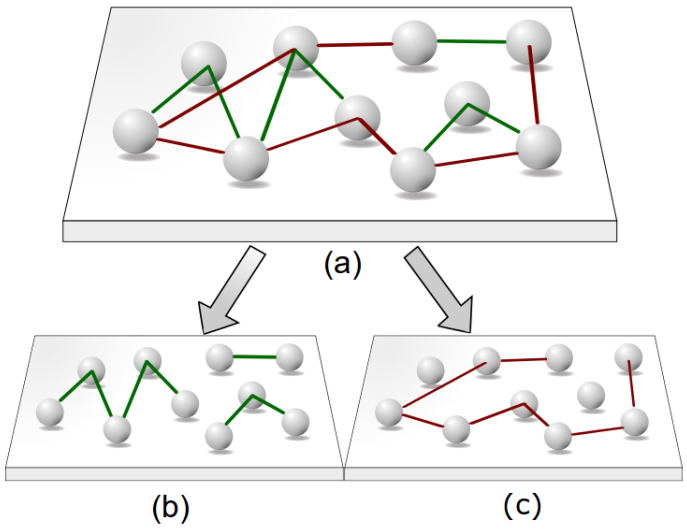

Figure 2.1: Extraction of the positive (b) and negative (c) subnetworks from a signed network (a).

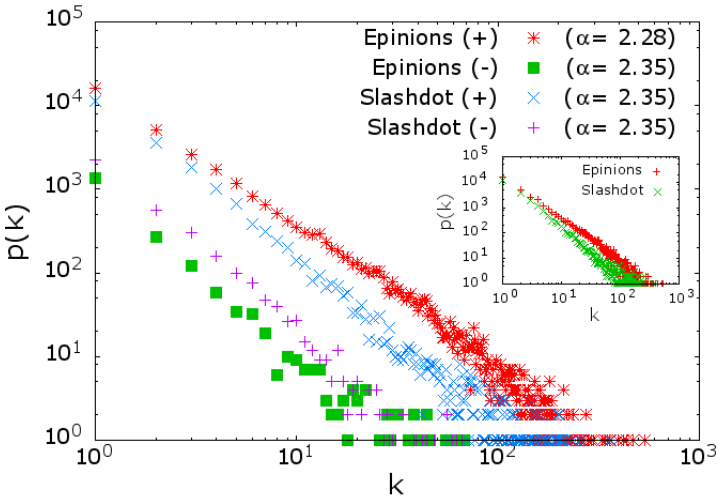

Figure 2.2: **Degree distributions of the Epinions positive ("trust") and negative ("distrust") subnetworks and of the Slashdot positive ("friend") and negative ("foe") subnetworks**. The inset shows the degree distributions of the Epinions and Slashdot unsigned networks with reciprocated links.

2.3 Degree correlations

Research has typically relied on two fundamental measures for detecting mixing patterns by degree in complex networks. The first measure is the quantity $K_{nn}(k)$, namely the average degree of the nearest neighbors of nodes with degree k, defined in [75] as

$$K_{nn}(k) = \sum_{k'} k' p(k'|k). \tag{2.1}$$

The transitional probability $p(k'|k)$ can be defined as the the conditional probability that a link emanating from a node of degree k is connected to a node of degree k'

$$p(k'|k) = \frac{E_{kk'}}{\sum_{k'} E_{kk'}} \equiv \frac{p(k,k')}{q(k)}, \tag{2.2}$$

where $E_{kk'}$ is the entry of the symmetric matrix E that measures the number of links between nodes of degree k and nodes of degree k' for $k \neq k'$, and two times that number for $k = k'$, $p(k,k')$ is the joint probability that a randomly chosen link connects two nodes of degrees k and k', $q(k)$ is the probability that a randomly chosen link is attached to a node with degree k

$$q(k) = \frac{kp(k)}{\langle k \rangle}, \tag{2.3}$$

$p(k)$ is the degree distribution of the network, i.e., the probability that a node chosen uniformly at random from the network has degree k, and $\langle k \rangle = \sum_k kp(k)$ is the average degree over the whole network.

In uncorrelated networks, the joint probability $p(k,k')$ factorizes and can be expressed in terms of the degree distribution, i.e., $p(k,k') = \frac{kk'}{\langle k \rangle^2} p(k)p(k')$, thus yielding

$$K_{nn}(k) = \sum_{k'} k' \frac{p(k,k')}{q(k)} = \frac{\langle k^2 \rangle}{\langle k \rangle}. \tag{2.4}$$

Thus, if there are no degree correlations, $K_{nn}(k)$ does not vary as a function of k: regardless of the degree a node has, its nearest neighbors have on average the same degree. By contrast, an increasing (decreasing) behavior of $K_{nn}(k)$ as a function of

k indicates that the network is assortative (disassortative) by degree: as the degree of a node increases, the degree of the node's nearest neighbors tends, on average, to increase (decrease).

The second method for detecting degree correlations relies upon the assortativity coefficient, a measure originally proposed by Newman [66] that is a suitably modified version of the standard Pearson correlation coefficient for measuring the correlation between the degrees of adjacent nodes in a network. Given a randomly chosen node that lies at the end of a randomly chosen link, one can define the excess degree of that node as the number of links incident upon the node other than the one along which the node was reached [76]. The excess degree of such node is distributed according to

$$e(k) = \frac{(k+1)p(k+1)}{\langle k \rangle}. \tag{2.5}$$

The assortativity coefficient for detecting mixing by degree can now be defined as

$$r = \frac{1}{\sigma_e^2} \sum_{kk'} kk'(e(k,k') - e(k)e(k')), \tag{2.6}$$

where $e(k,k')$ is the joint probability that a randomly chosen link in the network connects a node that has excess degree k with a node with excess degree k', $\sigma_e^2 = \sum_k k^2 e(k) - [\sum_k ke(k)]^2$ is the variance of the distribution $e(k)$, and $e(k)e(k')$ is the expected value of the quantity $e(k,k')$ in the case in which links are placed between nodes uniformly at random regardless of the degrees of the connected nodes. The values of r lie in the range $-1 \leq 0 \leq 1$, with $r = 1$ indicating perfect assortativity, $r = -1$ perfect disassortativity, and $r = 0$ lack of degree correlations [66]. Table 2.2 shows the values of the Pearson coefficient r of various social networks and other types of networks. As indicated by the Table, social networks exhibit a positive value of r, while technological and biological networks a negative one.

I begin the analysis of degree correlations by uncovering mixing patterns from the the unsigned Epinions and Slashdot networks with reciprocated links. Fig. 2.3 shows a positive trend for $K_{nn}(k)$, as was typically documented in social networks. To analyze degree correlations in the signed subnetworks, I measured and plotted $K_{nn}(k)$ for all four subnetworks. As shown in Fig. 2.4, two main distinct patterns can be detected. The positive subnetworks show the typical structural signature

Type	Network	Size n	Assortativity r
Social	physics co-authorship	52,909	0.363
	biology co-authorship	1,520,251	0.127
	mathematics co-authorship	253,339	0.120
	film actor collaboration	449,913	0.208
	inter-locking directorates	7,673	0.276
	e-mail address books	16,881	0.092
Technological	Internet	10,697	-0.189
	World-Wide Web	269,504	-0.067
	software dependencies	3,162	-0.016
Biological	protein-to-protein interactions	2,115	-0.156
	metabolic network	765	-0.240
	neural network	307	-0.226
	marine food web	134	-0.263
	freshwater food web	92	-0.326

Table 2.2: **Evidence of assortativity in unsigned or positive social networks** [63]

of social networks, namely the tendency of nodes to connect to other nodes with a similar degree (assortative mixing by degree). By contrast, the negative subnetworks display disassortative mixing by degree: high-degree nodes tend to be connected with low-degree ones.

This finding is further corroborated by the values obtained for the correlation coefficient r. These values are $r^{+}_{Ep} = 0.219$ and $r^{-}_{Ep} = -0.017$ for the positive and negative Epinions subnetworks, respectively, and $r^{+}_{Sl} = 0.162$ and $r^{-}_{Sl} = -0.114$ for the positive and negative Slashdot subnetworks, respectively.

These results are in qualitative agreement with, and generalize, a widely supported empirical regularity found in a variety of social networks: when links have a positive connotation, or can be assumed to have a positive one, they tend to connect nodes with similar degrees [66, 67]. However, findings also suggest that, when links have a negative connotation, they tend to connect nodes with dissimilar degrees. In Fig.2.5 is shown a subset of the Slashdot network. It is clear the structural difference between the positive (green) links and the negative (red) ones.

Combined, these two sets of results undercut one of the arguments that the literature has proposed to explain degree correlations in social networks [67]. This

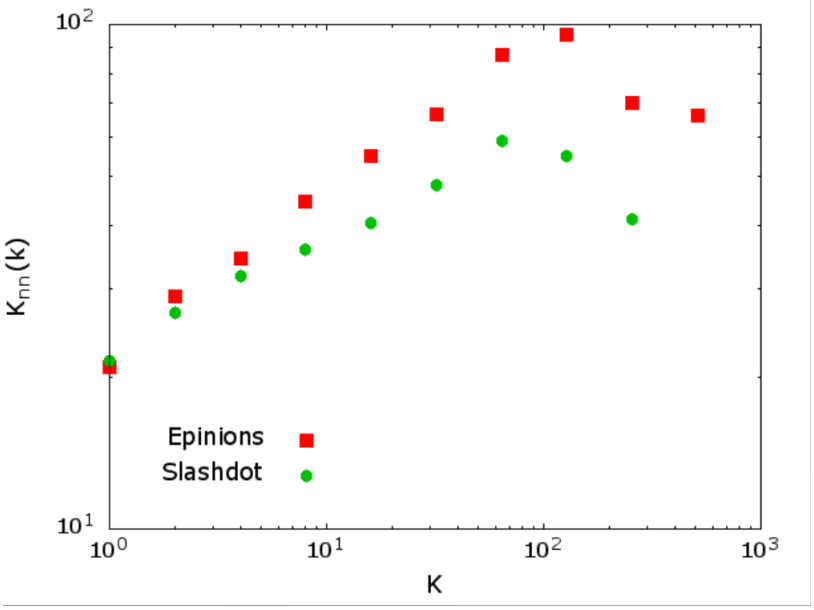

Figure 2.3: $K_{nn}(k)$ **for the Slashdot and Epinions unsigned networks with reciprocated links**. The observed positive trends are in qualitative agreement with the assortative patterns found in many other social networks.

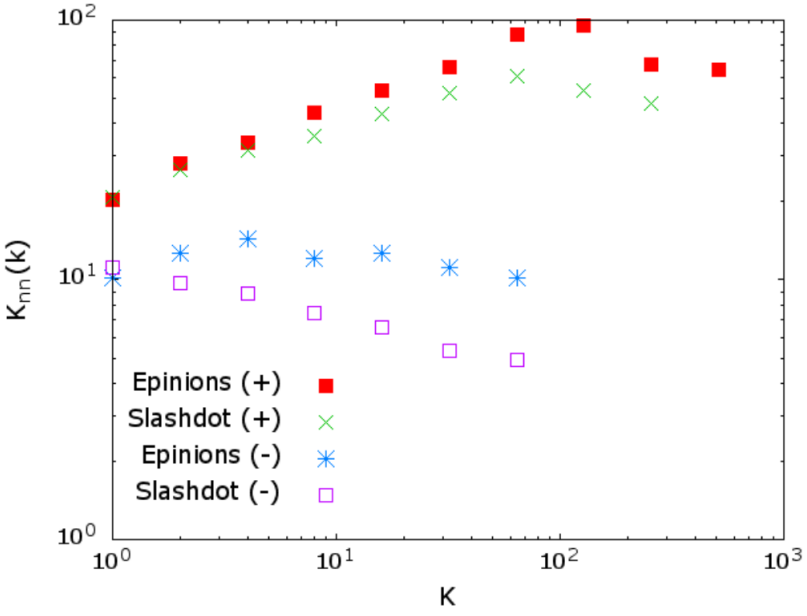

Figure 2.4: $K_{nn}(k)$ **for the Epinions and Slashdot positive and negative subnetworks**. The positive subnetworks display a positive trend, while the negative subnetworks display a negative trend.

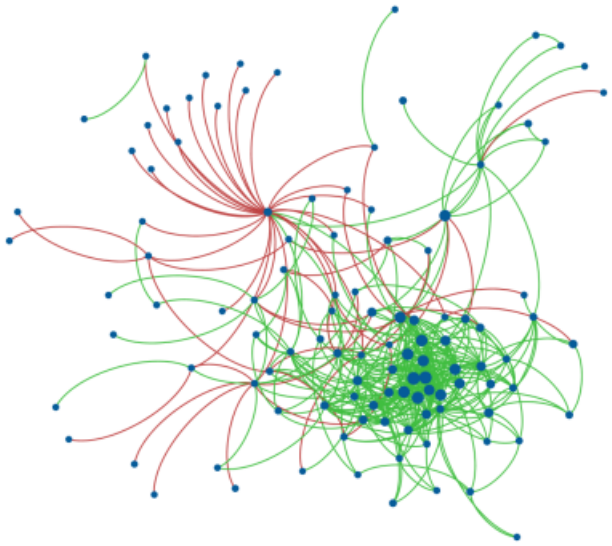

Figure 2.5: **A portion of Slashdots network**. Red links represent negative relationships between nodes, while green links positive ones. From this picture it is clear that the way which nodes connected is influenced by the nature of the relationships: nodes with many enemies are principally connected with nodes that have few enemies, while the opposite happens for the positive relationships.

argument is premised on the idea that assortative mixing is attributable to the tendency of nodes to coalesce into distinct communities. However, because this tendency can be detected in both the positive and negative subnetworks, community structure would in itself be not sufficient for explaining the assortative mixing patterns observed in the positive subnetworks. Other mechanisms are likely to be responsible for these patterns.

In both Epinions and Slashdot, individuals cluster into communities based on their common interests in the same products or news. However, the observed mixing patterns seem to originate not simply from common interests, but more precisely from the way individuals use the posted comments as cues for making positive or negative judgements on one another. More generally, the comparison between positive and negative subnetworks suggests that the observed degree correlations depend on the sign of the links between nodes. To gain a better understanding of this relation between sign of links and degree correlations, in what follows I will propose a class of simple generative models of signed networks.

2.4 Signed networks with degree correlations that depend on the sign of the links

I begin by focusing on signed random networks with binomial degree distributions, in which nodes can be split into two mutually exclusive groups. Subsequently, I will refine the analysis by investigating the case of assortative and disassortative signed networks with power-law degree distributions.

2.4.1 Signed random networks with binomial degree distributions

I draw on, and extend, a model originally developed by Newman and Park for undirected unsigned networks with multiple communities [67]. I create random networks with N nodes that satisfy the following requirements:

1. degrees are homogeneously distributed across the nodes;
2. each node can be a member of one of two mutually exclusive groups;
3. there are no degree correlations prior to the attribution of signs to the links; and
4. signs are associated with links in such a way that the resulting signed network is structurally balanced.

To obtain such networks, I apply the following rules:

1. any pair of nodes are connected through a link with a uniform probability p, and disconnected with probability $1 - p$;
2. given two groups, each node is assigned to one of them with probability m and to the other with probability $1 - m$; and
3. connections between nodes within the same group are associated with a positive sign, while connections between nodes from different groups with a negative sign.

A schematic representation of the polarization of a network into two distinct groups according to our model is shown in Fig.2.6. The model generates random uncorrelated networks with a binomial degree distribution. Notice that the original model proposed by Newman and Park corresponds to the case in which there is more than one community and $m = 1$ such that the resulting network is unsigned by construction. In our case, for the sake of simplicity, I introduced only one community as findings are qualitatively similar to those obtained with multiple communities. Moreover, as m approaches the value of 0.5, the network becomes perfectly polarized into two distinct groups of equal size. As m gets closer to either zero or one, polarization gradually disappears, and the network becomes increasingly dominated by one of the two groups [77].

Finally, to obtain a signed network, I attribute signs to links using an assignment rule that discriminates between links within and across groups. According to the *structure theorem* [36,78], the application of this rule of sign attribution ensures that the resulting signed network is structurally balanced. In accordance with the definition originally proposed by Heider [71] and subsequently extended by Cartwright and Harary [36], a network can be regarded as balanced when each of its cycles is positive, i.e., it includes an even number of negative links [79]. In turn, the structure theorem ensures that this is precisely the case when the network is polarized into two mutually exclusive subsets of nodes such that each positive link connects two nodes of the same subset and each negative link connects nodes from different subsets [36, 78].

As with the real networks, from the global signed network I extract two subnetworks, each including only positive or negative links. I then test whether and the extent to which network polarization has any critical role in the emergence of non-trivial mixing patterns in the positive and negative subnetworks. To this end, I simulate the model for an arbitrarily large value of N, and calculate the values of the correlation coefficient r between the degrees of connected nodes within the unsigned networks and the signed subnetworks obtained in correspondence of the different values of the probability m.

As indicated by Fig. 2.7, the positive subnetwork displays an assortative mixing by degree, as was observed in our two positive subnetworks as well as in many other social networks documented in the literature [66, 67]. By contrast, the negative

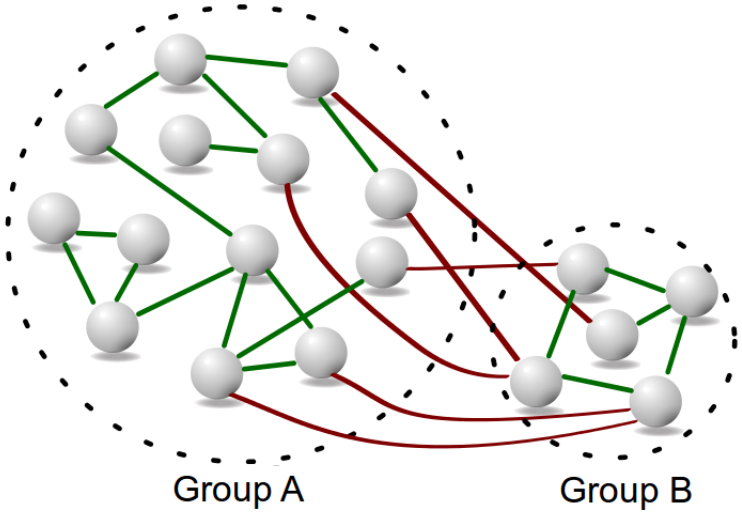

Figure 2.6: **Network polarization and sign attribution**. Schematic representation of the allocation of nodes into two mutually exclusive groups. Links between nodes belonging to the same group are positive (green), whereas links between nodes of different groups are negative (red).

subnetwork, like the ones extracted from both the Epinions and Slashdot networks, displays a disassortative mixing pattern. The sign of the links or, more precisely, the rules underpinning the attribution of sign to links, seem to be responsible for the variation in the mixing patterns. In particular, results suggest that non-trivial degree correlations of the signed networks would remain hidden and undetected if they were simply assumed to be the same as the ones of the corresponding unsigned networks obtained by removing or ignoring the signs of the links.

To further explore the conditions under which such degree correlations are likely to emerge in signed networks, in the subsequent section I will extend our analysis by using a number of more refined and realistic network generative models and by introducing additional combinations of structural properties of the networks. However, before I proceed in that direction, Inow formalize the properties of our current model in terms of the degree correlations displayed by the signed subnetworks. Given N nodes, and two mutually exclusive groups A and B, I set N_A to be the number of nodes that belong to group A, and $N_B = N - N_A$ the number of nodes that belong to group B. The probability that in group A there are N_A nodes can be expressed

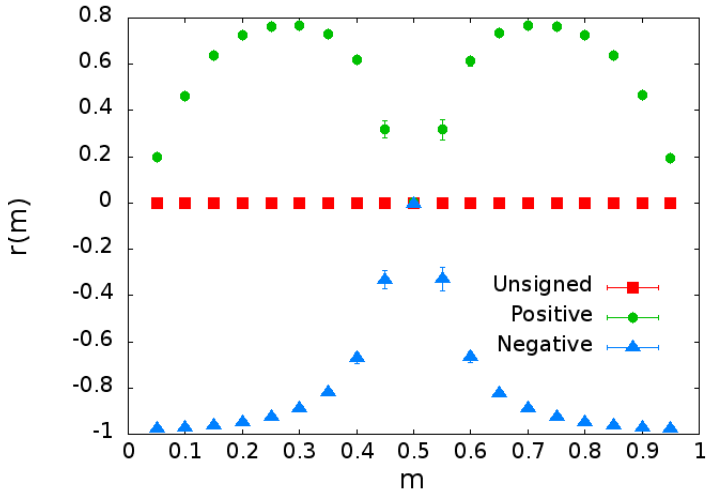

Figure 2.7: **Correlation coefficient r plotted against the probability m of being a member of one group**. The graph shows the trends of r for the positive and negative subnetworks obtained with the model, when $N = 10,000$ and $p = 0.01$. For each value of m, the correlation coefficient r is the average over 50 realizations of the network.

as

$$p(N_A) = \binom{N}{N_A} m^{N_A}(1-m)^{N-N_A}. \tag{2.7}$$

The "positive" degree $k_{A,+}$ of a node in group A is the number of positive links incident upon the node. The probability that a node that belongs to group A has a "positive" degree $k_{A,+}$ when the total number of nodes in group A is N_A, is given by

$$p(k_{A,+}|N_A) = \binom{N_A - 1}{k_{A,+}} p^{k_{A,+}}(1-p)^{N_A - k_{A,+}}. \tag{2.8}$$

In Eq.2.8, p represents the independent probability of a link in the network, and $k_{A,+}$ is the positive degree of nodes in group A (i.e., the number of links to other nodes in A). I define $K_{nn}^{A,+}(k_{A,+})$ as the average positive degree of the nearest friends of nodes of group A. Since, given a certain number of nodes in group A, the network formed by the links between these nodes is a random network (i.e., uncorrelated),

using Eq.2.4 I have

$$K_{nn}^{A,+}(k_{A,+}) = \sum_{N_A>0} p(N_A) \frac{\langle (k_{A,+})^2 | N_A \rangle}{\langle k_{A,+} | N_A \rangle} \simeq Npm, \tag{2.9}$$

where the average in Eq.2.9 is taken over the distribution $p(k_{A,+}|N_A)$ defined in Eq.2.8, $P(N_A)$ is defined in Eq.2.4.1, and where I have assumed $N \gg 1$.

I thus obtained a constant value for $K_{nn}^{A,+}(k_{A,+})$ that is independent of the positive degree $k_{A,+}$. In the same way, if I evaluate $K_{nn}^{B,+}(k_{B,+})$, i.e., the average positive degree of the nearest friends of nodes in group B, I obtain: $K_{nn}^{B,+}(k_{B,+}) = Np(1-m)$, which is also a constant function of $k_{B,+}$. As to the negative subnetwork, I obtain the same results for both groups of nodes. That is, $K_{nn}^{A,-}(k_{A,-}) = Npm$ is the average negative degree of the nearest enemies of nodes in group A, and $K_{nn}^{B,-}(k_{B,-}) = Np(1-m)$ is the average negative degree of the nearest enemies of nodes in group B. What differentiates the two groups in each signed subnetwork is simply the mean degree of their nodes. For instance, if nodes of group A (or B) in the positive subnetwork have an average positive degree of Npm (or $Np(1-m)$), in the negative subnetwork they have an average negative degree of $Np(1-m)$ (or Npm). In the case of $m = 0.5$, namely when groups are of equal size, it would not be possible to distinguish between the two subnetworks (see Fig.2.7), and I thus obtain the same results as in the case of the uncorrelated unsigned network.

As suggested by Eq.2.9, the polarization of a network with a binomial degree distribution into two groups of heterogeneous size generates two distinct values of $K_{nn}(k)$ for each of the two subnetworks, namely Nmp and $(1-m)Np$. In other words, the overall values of $K_{nn}(k)$ for each signed subnetwork result from the different (and complementary) contributions of the two groups in which, in turn, nodes have positive and negative degrees of different (and non-overlapping) values (see Fig.2.8a). For instance, in the case of the positive subnetwork, and when $m > 0.5$ and A is the larger group, the contribution to the overall $K_{nn}^{+}(k_{+})$ from group A is $K_{nn}^{+}(k_{(A,+)}) = Npm$ (which in turn corresponds to the larger values of k_{+}), while the contribution from group B is $K_{nn}^{+}(k_{(B,+)}) = Np(1-m)$ (which corresponds to the smaller values of k_{+}). As indicated by Fig.2.8b, when the two contributions are combined, $K_{nn}^{+}(k_{+})$ takes on two distinct constant values in correspondence of

two distinct sets of values of the positive degree, thus yielding the positive trend that signals the assortative mixing pattern of the positive subnetwork. Similarly, the negative trend of $K^-_{nn}(k_-)$ for the disassortative negative subnetwork results from the combination of the two distinct and complementary contributions from the two groups: $K^-_{nn}(k_{(A,-)}) = Npm$ from group A in correspondence of the smaller values of k_-, and $K^-_{nn}(k_{(B,-)}) = Np(1-m)$ from group B in correspondence of the larger values of k_-.

These trends are primarily due to the value of m which, in turn, affects the *opportunity* for nodes to create links within and across groups. Notice that, on average, the value of the degree of a randomly chosen node from a network with a standard binomial degree distribution is Np, regardless of which group the node belongs to. However, the polarization of the network into two groups of unequal size (i.e., $m \neq 0.5$), in combination with our rule of sign attribution, generates heterogeneity across nodes in terms of the proportion between positive and negative links incident upon them. Let us suppose that group A is the larger one. Each node, regardless of the group it belongs to, is surrounded approximately (for large N) by Nm potential neighbors from the dominant group (A) and $(1-m)N$ potential neighbors from the smaller group (B). Thus, each node, regardless of its affiliation, is likely to direct most of its links toward the nodes that belong to the larger group. This, in turn, has a direct bearing on the relative number of friends and enemies a node can have, depending on the group it belongs to. Because a node that belongs to the larger group has a higher chance than a node from the smaller group to direct links toward nodes of its own group (i.e., A), then as a result of our rule of sign attribution a node from the larger group also has a higher chance than a node form the smaller group to create friends by forging positive links with others. By contrast, a node from the smaller group (B) is more likely than a node from the larger group (A) to create links across groups, which in turn leads the former node also to be more likely to create more negative links than the latter. This difference in opportunity of "signed interactions" is responsible for the two different values obtained for the positive and negative $K_{nn}(k)$ attributable to the two groups of nodes, and can ultimately explain the assortative and disassortative mixing patterns, respectively of the positive and negative subnetworks.

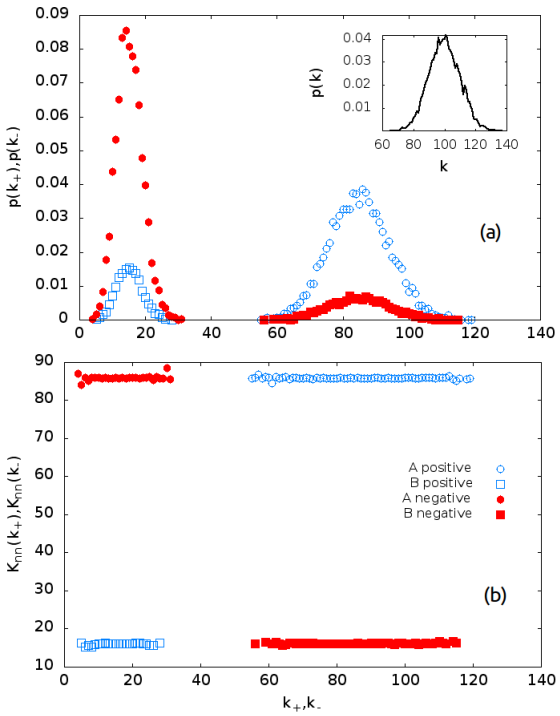

Figure 2.8: **The positive and negative degree distributions $p(k_+)$ and $p(k_-)$ and the trends of $K^+_{nn}(k_+)$ and $K^-_{nn}(k_-)$ for a network with a binomial unsigned degree distribution and polarization into two groups**. A network with a binomial degree distribution was created, with $N = 10,000$, $p = 0.01$ and $m = 0.85$, and in which A is the larger group. Panel (a) shows the positive and negative degree distributions, $p(k_+)$ and $p(k_-)$. Findings indicate the two distinct distributions for each signed subnetwork, one attributable to group A and the other to group B. The inset shows the degree distribution of the unsigned network. Panel (b) displays the trend of $K^+_{nn}(k_+)$ and $K^-_{nn}(k_-)$, respectively for the positive and negative subnetworks. For each subnetwork, the value of $K_{nn}(k)$ is constant within the same group, i.e., 85 for group A and 15 for group B. The two panels indicate that there are two corresponding gaps between values for $K_{nn}(k)$ and the signed degree distributions. This is due to the fact that the minimum value of degree in the unsigned network is 60 (see inset). Each node, regardless of the group it belongs to, has on average 85% of its neighbors from group A. If the node with degree 60 belongs to group A (B), it has, on average, $k_+ = 51$ ($k_- = 9$) in the positive (negative) subnetwork. Similarly, the maximum negative (positive) degree for a node in group A (B) would depend on the maximum value of the degree in the unsigned network, i.e., 140, yielding $k_- = 21$ ($k_+ = 119$). This therefore causes a gap between degrees ranging from 21 to 51, as shown in both panels. Panel (b) indicates the positive and negative trends for $K_{nn}(k)$, respectively for the positive and negative subnetworks, when both contributions from the two groups are taken into account.

2.4.2 Signed networks with power-law degree distributions

Previous empirical research has documented a large number of social networks characterized by statistically heterogenous connectivity: while the majority of nodes have only few connections, a minority have a disproportionally large amount of links to other nodes [14]. For this reason, I now move beyond the case of random networks with binomial degree distributions, and study the mixing patterns of more realistic signed networks characterized by power-law degree distributions. To this end, I introduce a generative model of scale-free signed networks. The choice of the model is also motivated by the need to ensure that the resulting unsigned network (i.e., the network obtained prior to the allocation of signs to links) is characterized by non-trivial degree correlations. This, in turn, will serve a two-fold purpose. First, it will help create networks with structural properties that are comparable to those observed in a variety of real-world networks [66, 75, 80]. Second, it will allow us to investigate whether the degree correlations of the unsigned network may be responsible for the difference between the mixing patterns of the positive and negative subnetworks.

I begin by constructing unsigned networks characterized by a power-law distribution and assortative mixing by degree. This will enable us to replicate the patterns observed in both the Slashdot and Epinions unsigned networks (see Fig.2.3). Among the models that satisfy the above requirements, in what follows I will use the copying model [81] and an extension of the rewiring model proposed by Xulvi-Brunet and Sokolov [82] based on the scale-free Barabási-Albert network [14].

First, the copying model begins with an initial connected network of n nodes. At each step, a new node is added to the network and another incumbent node is selected by chance: with a probability p the new node will create a link with one of the neighbors of the selected node, and with a probability $1 - p$ it will create a link with a node selected at random. Second, the rewiring model [82] is suitably applied to an initial network with a given scale-free degree distribution obtained by following the rules of the Barabási Albert model [14]. The rewiring process is then modeled as follows: (i) two links are selected at random; (ii) the four nodes connected through these two links are sorted in increasing order of degree; (iii) if the first two nodes and the last two nodes are not connected, links are rewired accordingly; otherwise,

(iv) the two links are dismissed, and a new pair of links are selected. After several iterations, an assortative network can be obtained. Both methods indeed generate an unsigned, undirected and assortative network characterized by a power-law degree distribution.

Drawing on these two generative models, I obtain unsigned networks that I then transform into signed networks by applying the last two rules from the basic model in Section 2.4, namely: (i) polarization of the network into two mutually exclusive groups of nodes; and (ii) attribution of a positive sign to links within groups and a negative sign to links across groups. Just as with the uncorrelated case, I then extract the positive and negative subnetworks from the signed networks, detect the mixing patterns of these subnetworks, and compare them with the patterns observed in the unsigned network.

To shed light on the role of the sign of links in the emergence of mixing patterns, I vary the rules governing network polarization and sign attribution, and extract and assess the corresponding signed subnetworks. First, as with the uncorrelated case, I manipulate network polarization by varying the degree to which the two groups differ in size. To this end, I use different values of m, the probability that a node belongs to one of two groups: as usual, at $m = 0.5$, the network is perfectly polarized, while for values approaching zero and one, the network becomes increasingly homogeneous and dominated by one single group [77].

Second, by manipulating our rule of sign attribution, I aim to vary the degree to which the signed network is structurally balanced. Previous research has long provided empirical evidence in favor of the tendency of individuals to avoid or alleviate cognitive tension by transforming an unbalanced structure into a balanced one [83, 84]. Yet, a number of studies have equally suggested that many observed signed structures for social groups are not structurally balanced, at least when they are assessed at single points in time [85, 86]. To account for such empirically documented variations in structural balance, in what follows I test whether this property, in combination with other conditions, is indeed necessary for the emergence of non-trivial mixing patterns within signed networks that differ from those observed in the corresponding unsigned networks. In this sense, I extend our previous analysis by investigating whether the sign of links can still produce some effect upon degree correlations also when the network is unbalanced.

Notice that, as implied by the structure theorem [36, 78], to obtain structurally unbalanced networks, it would not be possible to divide the population of nodes into two even or uneven groups and then impose our homophily-based rule of sign attribution (i.e, positive links within groups and negative links across groups). Indeed, from the structure theorem it follows that this procedure would necessarily generate a structurally balanced network. To obtain an unbalanced network, I therefore reshuffle the signs of the links within the corresponding balanced networks. In this way, the random reallocation of signs to links transforms the network from a balanced to an unbalanced state.

In summary, starting from assortative unsigned networks with a power-law degree distribution, I create four distinct groups of signed networks and corresponding subnetworks by combining the following structural conditions: (i) even versus uneven allocation of nodes into two mutually exclusive groups; and (ii) balanced versus unbalanced network structure. For the sake of simplicity, I label the four groups of networks as follows:

Ass/Het/Bal: (i) The unsigned network is assortative; (ii) nodes are heterogeneously allocated to groups; and (iii) the signed network is balanced.

Ass/Hom/Bal: (i) The unsigned network is assortative; (ii) nodes are homogeneously allocated to groups; and (iii) the signed network is balanced.

Ass/Het/Un: (i) The unsigned network is assortative; (ii) nodes are heterogeneously allocated to groups; and (iii) the signed network is unbalanced.

Ass/Hom/Un: (i) The unsigned network is assortative; (ii) nodes are homogeneously allocated to groups; and (iii) the signed network is unbalanced.

Results are shown by Fig.2.9, in which the unsigned assortative networks were generated through the copying model [81]. Each panel of Fig.2.9 shows the trends of $K_{nn}(k)$ for the unsigned network, for the positive subnetwork, and for the negative subnetwork obtained under each of the four combinations of structural conditions. Findings clearly indicate that most signed subnetworks retain the assortative pattern that characterizes their corresponding unsigned networks. There is, however, an exception: as indicated by panel (a) of Fig.2.9, there is one case in which a decreasing

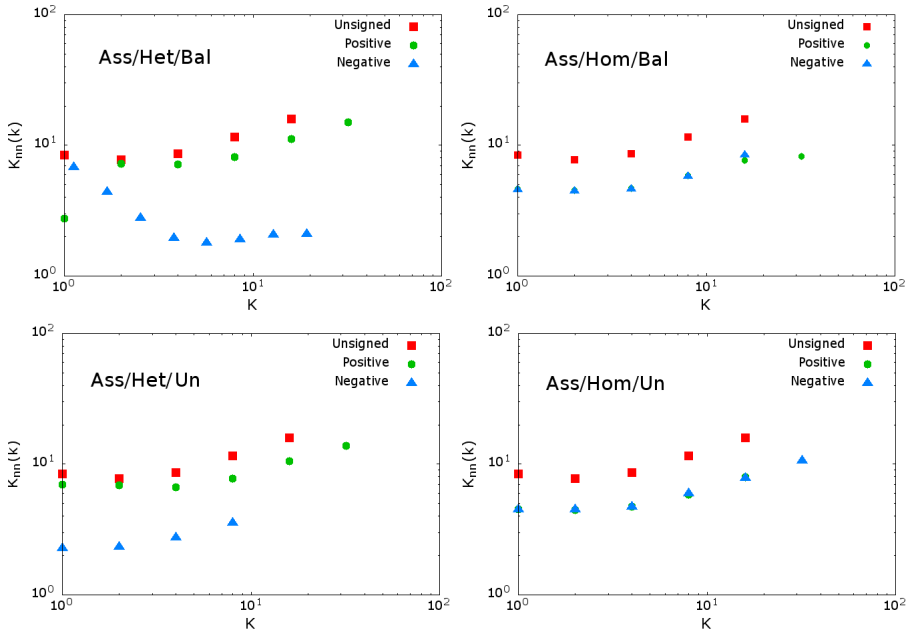

Figure 2.9: **Positive and negative subnetworks obtained from an assortative unsigned network with power-law degree distribution**. The unsigned network was generated through the copying model with $N = 10^4$ nodes. Findings indicate that different mixing patterns for the positive and negative subnetworks are obtained only when the assortativity of the unsigned network is combined with the heterogeneous allocation of nodes into groups and with the presence of structural balance.

trend of $K_{nn}(k)$ for the negative subnetwork is associated with an increasing trend for the unsigned network and the positive subnetwork. In particular, this opposite trend in mixing patters occurs when the following three conditions are jointly satisfied:

1. the unsigned network is assortative;
2. nodes are unevenly allocated into two mutually exclusive groups; and
3. the signed network is structurally balanced.

Under the above conditions, the disassortative pattern of the negative subnetwork would therefore remain hidden if the signs of links were removed from the global signed network and the nature and intensity of the mixing patters were simply inferred from the resulting unsigned network. Similar results are obtained when

the assortative unsigned network is created by applying the rewiring model by Xulvi-Brunet and Sokolov [82] to the scale-free Barabási-Albert network [14]. In this case, once again the negative subnetwork exhibits a variation in mixing patterns and becomes disassortative when the unsigned network is assortative, the groups are uneven in size, and the signed network is balanced.

I now test whether the mixing patterns in the positive and negative subnetworks differ when the unsigned network is disassortative. To this end, I create an unsigned network following the rules of the fitness model of growing networks, originally proposed by Bianconi and Barabási [87]. The results from our simulations are shown by Fig.2.10, in which the trend of $K_{nn}(k)$ is reported. If the unsigned network is characterized by a disassortative pattern, the patterns for the positive and negative subnetworks will always have the same trend across any of the four possible combinations of our two initial conditions. Subnetworks will always retain their disassortativity, regardless of the structural balance of the global network and the size of the groups.

Table 2.3 reports the correlation coefficient r of the degrees of connected nodes, for each of the networks and subnetworks analyzed above. The Table clearly indicates that there is only one case in which the mixing patterns of the positive and negative subnetworks differ. This variation indeed occurs when the unsigned network is assortative, the signed one is balanced, and groups differ in size. Under this combination of structural conditions, the correlation coefficient becomes negative for the negative subnetwork, while it remains positive for the positive one. Similar results are obtained when the assortative unsigned network is created by using the rewiring model by Xulvi-Brunet and Sokolov [82].

The reason for the opposite trends in the mixing patterns of the two signed subnetworks is similar to the one that explains the transformation of an unsigned uncorrelated random network into correlated signed subnetworks. As before, this reason is two-fold. First, the polarization of the network into two groups of unequal size is responsible for the heterogeneous distribution across nodes of opportunities of creating links within and across groups. Second, the requirement of structural balance (i.e., the rule of sign attribution) transforms these heterogeneous opportunities of social contact into equally heterogeneous opportunities to create friends or enemies. While a node of the larger group has a higher chance than a node

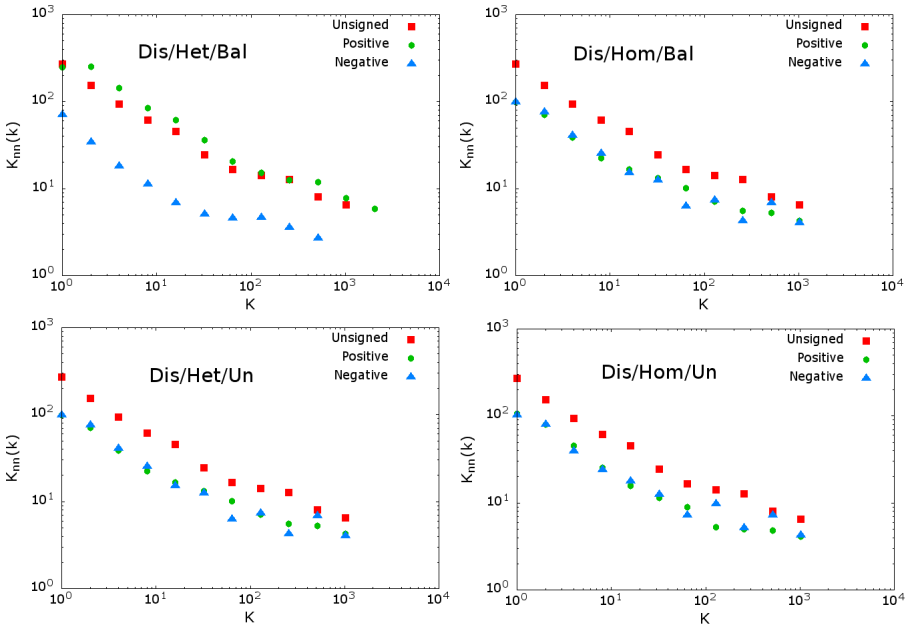

Figure 2.10: **Positive and negative subnetworks obtained from a disassortative unsigned network**. The unsigned network was obtained by removing the signs from the links of the network generated through the fitness model of growing networks, with $N = 10^4$ nodes. The unsigned network is characterized by a power-law degree distribution. Results indicate that across all combinations of the three conditions the mixing patterns for the positive and negative subnetworks have the same trend. In all panels a log-binning plot was reported in order to obtained a clearer trend for $K_{nn}(k)$.

The case of two groups		
Conditions	**Dis. Unsigned Network**	**Ass. Unsigned Network**
Het/Bal	$r^u = -0.09413$ $r^+ = -0.09853$ $r^- = -0.10864$	$r^u = 0.16249$ $r^+ = 0.15598$ $\mathbf{r^- = -0.3509}$
Hom/Bal	$r^u = -0.09413$ $r^+ = -0.09199$ $r^- = -0.09692$	$r^u = 0.16249$ $r^+ = 0.12062$ $r^- = 0.14804$
Het/Un	$r^u = -0.09413$ $r^+ = -0.093570$ $r^- = -0.096497$	$r^u = 0.16249$ $r^+ = 0.15243$ $r^- = 0.08244$
Hom/Un	$r^u = -0.09413$ $r^+ = -0.09748$ $r^- = -0.090807$	$r^u = 0.16249$ $r^+ = 0.12596$ $r^- = 0.13250$

Table 2.3: **Values of the correlation coefficient r for the case of polarization of the network into two groups.** The coefficient was calculated for each of the four combinations of structural balance (Bal) and unbalance (Un), and even (Hom) and uneven (Het) group size. Under each of the four combinations, the coefficient was calculated distinctively for each of the two cases in which the unsigned global network is assortative (and obtained through the copying model) and disassortative (and obtained with the fitness model). The variation in sign of the correlation coefficient between the positive and negative subnetworks occurs only when the unsigned network is assortative, the signed network is balanced, and groups are of unequal size.

of the smaller group to create intra-group connections, the latter node will have a higher chance to create inter-group connections than the latter. This imbalance of opportunities will be translated into the differential propensity nodes will have to create friends or enemies, depending on which group they belong to. It then follows that, when the whole unsigned network is assortative (disassortative), the positive subnetwork will remain assortative (disassortative) as it only includes intra-group connections between nodes of comparable propensity to make friends. Conversely, because the negative subnetwork only includes inter-group links, it will connect nodes that differ in their propensity to make enemies. For this reason, it will always remain disassortative, also when the unsigned network is assortative.

2.5 Extending the model: The case of three groups

Following the theoretical avenue that led Davis [88] to generalize the formalization of the theory of structural balance, I extend our model with network polarization to also account for the case in which nodes can be allocated to three or more mutually exclusive groups. As observed by Davis [88], individuals often split into more than two mutually hostile groups. To take this into account, Davis provided a generalization of the structure theorem [36, 78] by uncovering the necessary and sufficient condition for a signed network to be clusterable into two or more groups of nodes such that links connecting nodes within the same group are positive, and links connecting nodes from different groups are negative. Such condition was identified in the absence of cycles with exactly one negative link. It follows that all structurally balanced networks are clusterable, but not vice versa. Whether clusterable networks are also balanced depends on the number of disjoint groups of nodes.

The analysis carried out by Davis provides us with a theoretical backdrop against which I can further refine our model. First, I investigate whether our model is robust against the number of groups, namely whether the same results are obtained when the network splits into more than two mutually exclusive groups, but still remains structurally balanced. Second, I study our model in the more general case in which the network is clusterable into more than two groups, but it is not balanced. In what follows, I will focus our attention only on the case of three groups. The analysis can easily be generalized to any number of mutually exclusive groups.

Fig.2.11 shows a schematic representation of a network that splits into three mutually exclusive groups. The rule of sign allocation remains the same as before: links between nodes of the same group are assumed to be positive, and links between nodes from different groups negative. Let us assume that each node can belong to one of the three groups with a given probability p. I then have four possible cases:

1. homogeneous allocation of nodes into groups of equal size, i.e., $p_1 = p_2 = p_3$;
2. heterogeneous allocation of nodes into groups of uneven size, such that one group dominates the other two, i.e., $p_1 > p_2 \simeq p_3$;
3. heterogeneous allocation of nodes into groups of uneven size, such that two equally sized groups dominate a less populated one, i.e., $p_1 \simeq p_2 > p_3$; and

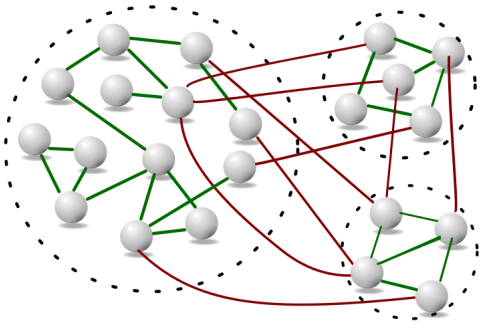

Figure 2.11: **The case of three mutually exclusive groups**. Schematic representation of the allocation of nodes into three groups such that links connecting nodes of the same group are positive (green), and links between nodes from different groups are negative (red).

4. heterogeneous allocation of nodes into groups of uneven size, such that, for any two groups, one dominates the other, i.e., $p_1 > p_2 > p_3$.

In what follows, I will concentrate on the first two cases. Results concerned with the third case will not be reported here because they are qualitatively similar to what is obtained with: (i) two equally sized groups, when the two dominant groups are much larger than the third one; and (ii) three equally sized groups, when differences in size become negligible. Similarly, the fourth case can be reduced to the previous cases, depending on the difference in size between groups.

To create a structurally balanced network, I impose the following constraint. When there is a (negative) link between two nodes that belong to two different groups, the two connected nodes are not allowed to share a common enemy, that is they are not allowed to be connected with the same node from the third group. In this case, each of the two nodes will change the target of the link to the third group, so as to avoid triangles in which all links are negative. In other words, two nodes may share a common enemy either when they are not connected themselves, or when they are connected and belong to the same group. In this sense, allowing coalition formation against a common enemy to occur only between nodes of the same group will preserve our rule of sign allocation that confines positive links only within, but not across, groups.

The trend of $K_{nn}(k)$ for the case of three groups is similar to the one obtained

with two groups. Fig.2.12 reports the value of $K_{nn}(k)$ for the unsigned network and signed subnetworks under the joint conditions of assortative unsigned network, structural balance, and uneven allocation of nodes into three groups (i.e., condition 2 above). As was the case with the two groups, the negative subnetwork, unlike the positive one, is characterized by a disassortative mixing pattern. As before, these opposite trends in mixing patterns do not emerge under all the other combinations of conditions, and in particular when networks are clusterable yet unbalanced [88]. Results thus suggest that clusterability is not a substitute for balance: networks that contain all-negative triangles connecting nodes from distinct groups do not display correlation patterns that differ from those obtained from unbalanced networks.

Table 2.4 further corroborates the results from Fig.2.12. The Table reports the values of the correlation coefficient r of the degrees of connected nodes in the unsigned network and the signed subnetworks. Just as in the case of two groups, the mixing pattern of the negative subnetwork differs from the patterns of the unsigned network and positive subnetwork only when the unsigned network is assortative, the signed network is balanced, and the three groups differ in size. Indeed under these conditions, the correlation coefficient is negative for the negative subnetwork, while it remains positive for the positive one.

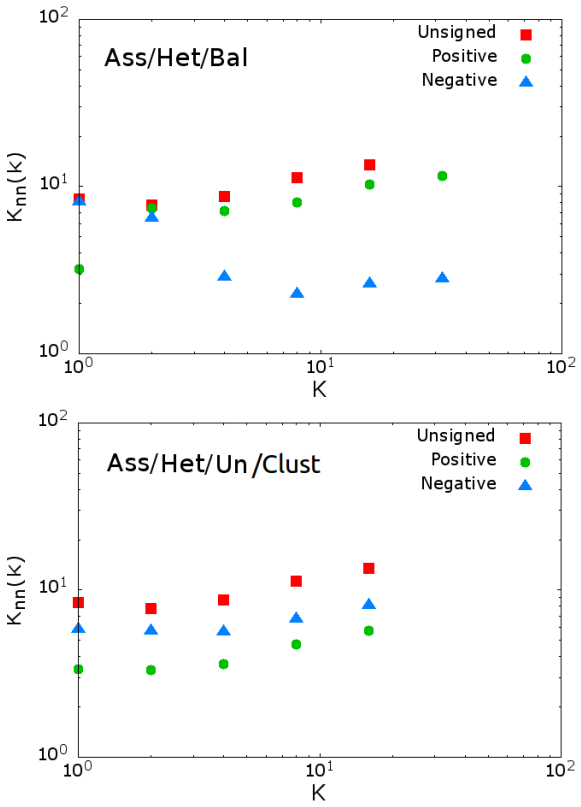

Figure 2.12: **Positive and negative subnetworks obtained from an assortative unsigned network in the case of three groups**. The unsigned assortative network was obtained as in Fig.2.9. Panel (a) reports the different mixing patterns for the positive and negative subnetworks obtained under the conditions of structural balance and heterogeneous allocation of nodes into three groups. Panel (b) reports results obtained for a network with groups of unequal size and that is clusterable (Clust) but unbalanced.

The case of three groups		
Conditions	**Dis. Unsigned Network**	**Ass. Unsigned Network**
Het/Bal	$r^u = -0.09413$ $r^+ = -0.08933$ $r^- = -0.11434$	$r^u = 0.16249$ $r^+ = 0.19226$ $\mathbf{r^- = -0.2404}$
Hom/Bal	$r^u = -0.09413$ $r^+ = -0.09234$ $r^- = -0.09535$	$r^u = 0.16249$ $r^+ = 0.15685$ $r^- = 0.14271$
Het/Un	$r^u = -0.09413$ $r^+ = -0.092480$ $r^- = -0.095247$	$r^u = 0.16249$ $r^+ = 0.15243$ $r^- = 0.07686$
Het/Un/Clust	$r^u = -0.09413$ $r^+ = -0.08364$ $r^- = -0.09524$	$r^u = 0.16249$ $r^+ = 0.14912$ $r^- = 0.12176$
Hom/Un	$r^u = -0.09413$ $r^+ = -0.09338$ $r^- = -0.09395$	$r^u = 0.16249$ $r^+ = 0.13252$ $r^- = 0.12230$
Hom/Un/Clust	$r^u = -0.09413$ $r^+ = -0.093371$ $r^- = -0.094508$	$r^u = 0.16249$ $r^+ = 0.18380$ $r^- = 0.10900$

Table 2.4: **Values of the correlation coefficient r for the case of a network that splits into three groups.** The coefficient was calculated for each of the four combinations of structural balance (Bal) and unbalance (Un), and even (Hom) and uneven (Het) group size. Under each of the four combinations, the coefficient was calculated distinctively for each of the two cases in which the unsigned global network is assortative (and obtained through the copying model) and disassortative (and obtained with the fitness model). The variation in sign of the correlation coefficient between positive and negative subnetworks occurs only when the unsigned network is assortative, the signed network is balanced, and the three groups are of unequal size such that one dominates the other two. The coefficient was also evaluated for the cases in which the network is clusterable (Clust) but unbalanced. Results are in qualitative agreement with the values obtained when the network is unbalanced and unclusterable.

2.6 Conclusions

This study was prompted by the empirical analysis of two signed social networks and by the observation that their mixing patterns by degree vary depending on the sign of the link. In particular, findings indicates that negative subnetworks are characterized by disassortative patterns, in sharp contrast with their corresponding unsigned networks and the positive subnetworks. The emergence of opposite trends of mixing patterns seems to be at variance with the widely accepted belief that social networks are predominantly assortative, possibly as a result of their tendency to organize themselves into communities [67]. Because both the positive and negative subnetworks have an underlying community structure, it follows that the social nature of links and the partition of nodes into communities are not, in themselves, a sufficient reason that explains why some observed social networks exhibit positive degree correlations. Results indeed seem to suggest that the pattern of such correlations depends on the sign of the links between nodes, and thus ultimately on the type of the social relationship between individuals.

To study the relation between sign of links and mixing patterns, I proposed a class of simple models in which nodes split into two mutually exclusive groups. I show the simple case of unsigned random uncorrelated networks, and then extended the analysis by also investigating unsigned correlated networks with power-law degree distributions, and cases in which the network is organized into three or more groups. Upon attribution of signs to the links of an originally unsigned network, two distinct signed subnetworks could be extracted, each including only links with a positive or negative sign. The comparative assessment of the degree correlations in these subnetworks suggested that, when the signed network is structurally balanced and the groups differ in size, the negative subnetwork is *always* characterized by a disassortative pattern, regardless of the correlation patterns displayed by the positive subnetwork and the corresponding unsigned network. In particular, under the combined conditions of structural balance and uneven group size, the correlation patterns of the two signed subnetworks differ when the unsigned network is either uncorrelated or assortative. In either case, the positive subnetwork is assortative, while the negative one is disassortative. In particular, the case of networks that split into three or more mutually exclusive groups suggested that clusterability is not a

substitute for balance: when networks are clusterable but unbalanced, both signed subnetworks display the same degree correlations as the one in the corresponding unsigned network.

By identifying the conditions under which degree correlations vary depending on the sign of the links, this study suggests that ignoring the sign would result in a loss of information on the structural properties of the network that would simply remain hidden in the unsigned network. Moreover, findings indicate that assortativity, often regarded as a characteristic signature of most social networks, can be justified not simply by the social character of these networks, but more precisely by the positive nature of the social relationships they embody. Indeed the broad category of social networks typically subsumes a variety of relationships and interactions that are often difficult to disambiguate and may, as result, intermingle with each other and remain confounded in one single type of connection. In such cases, detecting assortativity in a network may simply indicate either that the nature of the social relationships was ignored or that their positive components outweigh the negative ones. Conversely, disassortativity may indicate that the unsigned network is in itself disassortative or that the negative components of the links outweigh the positive ones. Finally, a lack of degree correlations may result simply from an unsigned uncorrelated network or from cases in which the positive and negative components of the relationships compensate each other out.

This analysis can be regarded as a platform for further studies of mixing patterns in complex networks. If degree correlations vary according to the sign and nature of the connections, this study suggests that the sign of the links could, in principle, be inferred simply from the analysis of the structural properties of a network. From this perspective, findings can help inspire the development of a quantitative measure for uncovering the hidden sign of the links from the type of mixing patterns exhibited by a network. This would prove to be useful especially in cases where the sign of links could not be assessed directly or it would be too costly to do so. For instance, gauging the collaborative or competitive properties of the relationships within and between organizations is typically constrained by a number of biases originating from the subjective, multiplex and complex nature of such relationships. These biases, however, can easily be overcome when the sign and nature of the relationships can be extracted directly from the degree correlations of the intra- and inter-organizational

networks.

Chapter 3

The nature of competition in start-up ecosystems

> "*There is a book yearning to come out of me: about how we can build the new collaboration economy, and the role of 'openness' in our quest for efficient use of resources and as a driver of innovation*"
>
> — Robin Chase, CEO of Zipcar

> "*I have been up against tough competition all my life. I wouldn't know how to get along without it.*"
>
> — Walt Disney

In the previous chapter I have described how positive and negative connections among individuals in social networks can shape the network topology. However, positive and negative connections are not a prerogative of social networks. Other networks such as organization networks, in which nodes represent groups of people or companies, may exhibit positive (e.g. collaborative) and/or negative (e.g. competitive) relationships as well.

In this chapter, I will focus on one particular type of organization network whose nodes represent start-ups, i.e., early-stage and innovative companies. During the last

decade we have witnessed an unprecedented growth in the interest of entrepreneurs and governments in start-ups. Because of their large impact on the world's economy and society, studies related to start-ups have started to attract both the interest of scholars, who aim to understand the mechanisms that lead companies to success, and of investors, who see the opportunity to further their returns.

With the spread of start-ups around the world, websites such as CrunchBase.com and Angel.co have started to collect information on start-ups with the intention to discover industry and investment trends. Based on CrunchBase's website, I have constructed a database that contains information concerning the individuals that are (or have been) working in each start-up and the location of the companies' headquarters. This dataset will allow me to create different types of networks of start-ups, to produce geographic analyses, and to study the movement of employees from one company to another.

Moreover, each company registered on CrunchBase can declare one or more industry sectors (or markets) in which its business is involved. Unfortunately, this information is unstructured and often imprecise. In order to make use of the information related to market sectors, I will propose a method to group industry sectors into different categories. The method rests on the construction of the network of markets based on the co-occurrence of markets tags in each start- up. In order to extract a hierarchical structure out of this network I will combine two network analysis methodologies. As a result, I will create what I have defined as the *market macro categories*. I will make use of these categories to study the similarities and differences between nations (in which the start-ups' headquarters are located) through the use of a hierarchical clustering methodology.

The CrunchBase dataset also provides a list of companies that each start-up declares as its direct competitors. I will evaluate the effects of competition on both the mobility of employees between companies and on the success of the nation in which they are located. In order to do so, in both cases, I will use a network approach. Namely, I will use the information concerning the companies' competitors and their employees to create two start-up networks. In the first network (the *declared-competition*) a start-up i is connected to a start-up j if i considers j as its competitor. In the second network (the *mobility* network) two start-ups are connected if there has previously been an exchange of employees between them.

In order to evaluate the effects of competition on the mobility of employees between companies, I will produce a third network as the overlap between nodes and links among the declared-competition network and the mobility network. Results indicate that the number of overlapping links between these two networks in respect of the number of overlapping nodes is very small, suggesting that the presence of competition has a negative impact on the flow of employees between competitors.

Finally, I will move on to the study of the effects of competition on the success of the ecosystem of a nation, i.e., the set of start-ups whose headquarters are located in a nation. I will propose to quantify the success of a nation as the ratio between the number of start-ups that have either undergone an IPO, been acquired, and/or acquired other companies, and the total number of companies that reside in the same nation. In order to quantify the level of competition of a nation, I will define a novel measure that I have named as the *competition blocking* coefficient. Results indicate that the success of a nation anti-correlate with the presence of competition, suggesting that the more a national ecosystem is competitive, the less successful it is.

3.1 Introduction

3.1.1 Positive connections: collaboration and creativity among organizations, a short review

I start this chapter with a brief digression concerning positive connections among organizations. One of the best examples of combining the use of network analysis and positive connections between organizations is the work of Brian Uzzi and Jarrett Spiro, titled "*Collaboration and creativity: The small world problem*" [89] in which they analysed the effects of collaboration among artists on the "creativity" of Broadway musicals between 1945 and 1989.

Creativity has long been studied in various fields across the social, behavioural, and organizational sciences, both at the individual and the organization or team level [90–92]. Social contexts are the environment within which creativity can benefit from collaborations among different people or teams. Recently, there has been an increase in interest in network perspectives on innovation in domains typically

related to knowledge creation, and particularly scientific collaboration [76, 93, 94].

Collaboration across different domains or groups improves creativity as a result of the matching and sharing of diverse ideas, creative materials, and ways of thinking or behaving [47, 48]. This relates to the network foundations of social capital[1] and its effects on performance [1, 49, 95]. Over the last years, social network analysts have become interested in the network foundations of the arts and creativity [96, 97].

By creating and analysing a bipartite graph in which artists are connected to their affiliations, Uzzi and Spiro found that the generated network has the properties of a so called "small world" network, i.e., a network with a high clustering coefficient and a small shortest path length (the average shortest path length grows as the logarithm of the number of nodes). They realised that the presence of a small world network has a significant impact on creativity and on performance (measured in terms of financial and artistic success). In particular, they found that collaboration has a non-linear effect on performance: the more the network exhibits collaboration, the more performance benefits from it, but only up to a given threshold above which the effects reverse.

Uzzi and Spiro argue that the presence of a small world network allows creative ideas that are generated by teams (i.e., highly connected clusters) to diffuse towards other teams, and to produce different and original material. However, if the whole network becomes too connected, the set of ideas from which every team can draw on (through their connections) becomes the same, and the novelty generated by the recombination of other teams' ideas ends up to be common material for all network actors.

3.1.2 Negative connections: competition among organizations

Over the last decades, the way of running a small business has drastically changed with the raise of start-ups all around the world. In the collective imagination there is a myth that with a garage and the right idea, it is easy to make a billion dollars company. Stories about the birth of Facebook, Apple, and Microsoft, where a geek, a computer, and his or her idea were the starting point of a successful business,

[1] For a discussion on social capital, see Chapter 5.

have reinforced the belief that these factors are the key ingredients behind success. However, reality is often different. What is frequently omitted or rarely highlighted in these stories is that the environment surrounding these "geniuses" and the network of professional relationships that they constructed were fundamental in fostering and nurturing their success. For instance, Steve Jobs was undoubtedly an innovator and a visionary, but without the help of Stephen Wozniak, who was working at Hewlett-Packard (HP) while they were creating the first Macintosh prototype, and was able to bring pieces of hardware from HP to their garage, he would probably have had difficulties creating his innovative computer, and today we might not be surrounded by back-lighted bitten apples.

In a working paper, that I will not discuss in this thesis, my collaborators [2] and I found that a key factor that leads start-ups to success is a "good" network position. We propose a mathematical quantitative framework to define a good position and to predict the success of a start-up by correlating its success with its position in the network of professional relationships mediated by the social interactions among individuals (e.g., inventors, employees, advisors, or founders). Thus the connections created between start-ups can be interpreted as the transfer, recombination, and exchange of knowledge or know-how between them.

Our findings indicate that the success of a start-up can be predicted based on the company's structural position in the network. The success rate of our prediction method ranges between 30% and 50%, thus well above the rate typically achieved by private investors (i.e., 10%[3]) through costly and labour-intensive screening processes. We also find that the success of the prediction correlates with historical economic trends and downturns. In order to define the success of a start-up we assign a binary success variable to each start-up if they have either done an IPO, acquired another company, or have been acquired by another company. If at least one of these conditions is fulfilled, the start-up is considered successful.

Thanks to the availability of big data on start-ups and their employees, I am able to extract and define two other start-up networks on a worldwide scale. The first network I construct focuses on the relationships of competition between start-ups.

[2] Moreno Bonaventura, Pietro Panzarasa, and Vito Latora

[3] 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. (ref: crunchbase.com)

The nature of competition among two companies is either direct or indirect. For a start-up, direct competitors represent a higher obstacle than indirect ones. Direct competitors derive from the same industry sectors, producing or selling the same services or products. In order to survive the race of competition, new born companies are forced to create either innovative products or radically new industry sectors. I will create what I defined as the *declared-competition* network based on the list of companies that each start-up declares as their direct competitors. The second start-up network draws on the movement of people from one company to another, which I will define as the start-up *mobility network*. Combining the declared-competition and the mobility networks I will show that competition strongly prevents the movement of people between companies.

I then move to the analysis of innovative ecosystems, as the collection of start-ups whose headquarters are based in a given nation. To define an innovative ecosystem, the choice of a nation (or State in the case of the US) is an adequate unit of analysis as it imposes physical boundaries and makes it possible to obtain a good statistics in terms of the number of start-ups located in them. The flourishing of an innovative ecosystem such as Silicon Valley has been, for many years, linked to a culture of openness towards the free circulation of people between companies. I will propose a quantitative, rigorous framework to correlate the presence of competition between start-ups within national boundaries with the success of the overall set of companies located within the national boundaries.

Before I start the analysis of the start-up networks, I will describe the dataset from which I have retrieved the information on start-ups. Subsequently, I will show the different industry sectors associated with each start-up. Unfortunately, the information provided by CrunchBase is unstructured and often imprecise, which brings to the creation of spurious market associations to each company. I will propose a method to extract a hierarchical structure which associates each market to a macro category. Based on these macro categories, I will propose a methodology to characterise innovative national ecosystems and to outline their differences and similarities through a hierarchical clustering technique.

3.2 The data

The online publisher of technology industry news *TechCrunch* in 2007 created the start-up database CrunchBase (*www.crunchbase.com*). CrunchBase is considered "*a minor miracle[...], a kind of stats based Wikipedia for start-ups*" [4]. The database contains information about start-ups and the people who currently work or previously worked for them. Most of the information is manually managed by several contributors affiliated with the CrunchBase platform, and voluntarily by founders and investors. I collected the data from the Crunchbase.com Web API and subsequently stored them in a *Neo4J* database. Data in the local database is as at August 2016.

Geography - CrunchBase collects information about companies from all over the world. For each company I have information concerning where its headquarters are based. Locations are hierarchically structured, with information concerning the city and the nation in which a start-up's headquarters is located. A first analysis shows that 55% of the registered companies are located in the U.S. The old continent is the second bigger cluster in terms of the number of start-ups per nation led by the United Kingdom and Germany. Another highly dense geographical cluster is Israel, which, to date, is among the most prosperous and highly innovative start-up centre in the world. Less dense concentrations of start-ups can be found in Australia, East Asia, and South America, showing that the rising up of start-ups has become a worldwide trend. The growth and spread of start-up ecosystems around the world may appear similar: the trend for cities and nations is to economically invest start-ups to help them to develop and grow while hoping that they will become successful. But not all the national ecosystems are equal. In the next section I will propose a method to evaluate similarities and differences among countries.

Mobility - Each start-up present a list of employees who are playing (or have played) a role in it on their CrunchBase web-page. Each person is provided with a profile page (similar to a Linkedin user page) in CrunchBase in which it is possible to retrieve the list of companies that he or she is (or has been) involved in. Most of this information is time-stamped. This allows a reconstruction of historical movements of

[4] http://www.forbes.com/sites/edmundingham/2014/11/05/crunchbase-is-such-a-valuable-start-up-analysis-tool-but-the-problem-is-it-has-no-filter/#69b53be833c8

people among start-ups to be produced. The data is sometime incomplete or presents inconsistencies, such as a person's role starting at a date prior to the company's foundation. In this case, I have removed the inconsistencies, preserving only the most reliable information according to the trust-code value provided by CrunchBase itself, which evaluates the "goodness" of the information provided. I will use of the temporal information to describe the displacement of people from one company to another, with which I will create the start-up mobility network (see Section 3.4). Positions covered by the employees registered in CrunchBase spread on different areas and levels. The most frequent job titles that appear in the dataset are reported in Tab.3.1, and a pie chart representation is reported in Fig.3.1. Most of the people registered in CrunchBase hold the title of Chief Executive Officer (CEO), founder, co-founder, and/or vice-president. The job position rank is generated by aggregating all positions covered by all people along their career.

Markets - Further information related to each start-up that can be retrieved from CrunchBase data is the company's industry sectors or markets. Each company can indicate one or more industry sector through a free-text attribution method. Unfortunately this approach may produce misspelled sectors and imprecisions which, by consequence, generates the presence and the proliferation of spurious market attributions. However, it is still possible to extract some meaningful insight from the raw data. In fact, Tab.3.1 shows most used "market-tags" and in Fig.3.1 follows a pie chart representation. The possibility to freely add new market-tags does not allow CrunchBase to create a hierarchical classification of the markets. In the following section I will construct the network of markets based on the co-occurrence of market tags in each start-up. Through a combination of two network techniques I will show the presence of a hierarchical structure with which I define market *macro categories* to which each market will be assigned.

News - CrunchBase information is not just manually curated, it is also enriched by bots which daily scrape the web looking for news about IPOs, acquisitions, and funding rounds on other platforms. I am going to use this information in order to assign a success binary variable to each start-up that has been acquired, and/or has acquired other companies, and/or has undergone and IPO. Finally, I will use this success variable in order to evaluate the success of a whole nation.

Competition - CrunchBase does not only provide information related to the

Job title	N. of entries	Market sector	N. of companies
CEO	48104	Software	23678
Founder	38816	Mobile	8936
Co-Founder	36428	E-Commerce	5692
Vice-President	29477	Curated web	4527
Board of Directors	19420	Healthcare	4071
Board Member	14591	Advertising	3876
President	13146	Enterprise Software	2769
CTO	9060	Education	2357
CFO	6295	Services	2030
Advisor	8118	Biotechnology	2007
COO	7923	Consulting	1931
Director	6857	Finance	1865
Chairman	3667	Social media	1599
Associate	2510	Information Technology	1459
Principal	2318	Games	1389
Consultant	2306	Manufacturing	1370
Software Engineer	2208	Financial Services	1364
Investor	1486	Analytics	1072

Table 3.1: Top job titles and the most used market tags in Crunchbase.

location, employees, and industry sectors of a company, but it also explicates the relationship of competition between companies. Looking at the dictionary definition, two companies either belonging to the same (or similar) industry, or offering similar products or services are defined as competitors. Crunchbase goes beyond the usual definition of competition. In fact, competitors are not preassigned by the co-occurrence of two companies in the same industry. On the contrary, companies are free to explicitly indicate a list of other companies that they consider as their direct competitors. Through the use of these competitive relationships, I can define and study the start-up network of declared competitors with a worldwide coverage.

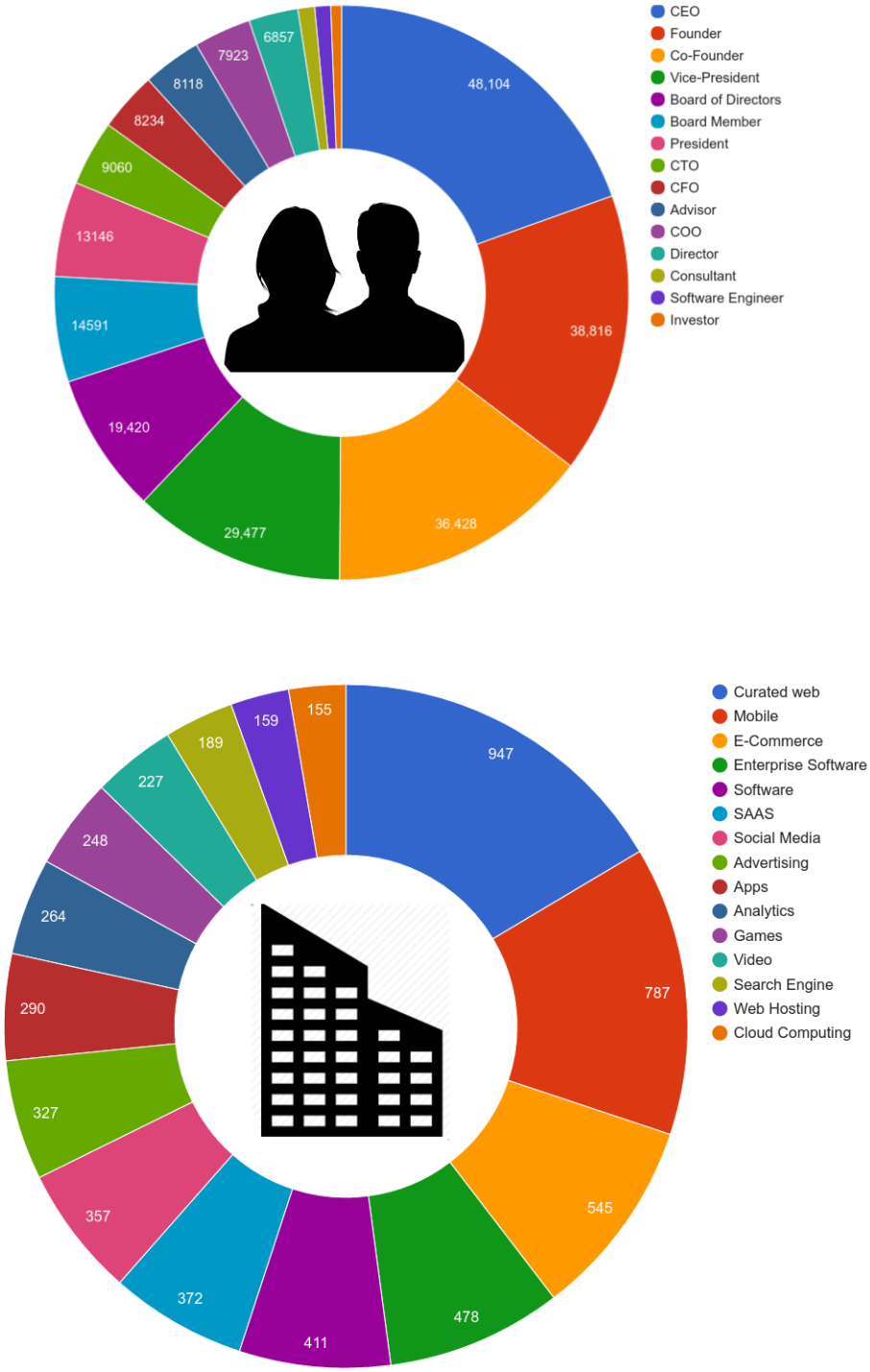

Figure 3.1: **Top job titles and industry sectors. (Top Figure)** Most common positions in the dataset. **(Bottom Figure)** Most common industry sectors in the dataset.

3.3 Extracting the market macro categories

Data on markets is not provided in a hierarchical structure, which makes it impractical to use it a as it is provided. I therefore apply the combination of two network-based methodologies in order to extract a hierarchy out of the data, trough creating market macro categories and assigning each single market to one of them.

3.3.1 The market network

First, I define as $\mathcal{M}$ the set of all markets in the dataset. Each start-up s is associated with a subset of markets $M_s \subseteq \mathcal{M}$ such that $M_s = \{m_1, m_2, ..., m_n | m \in \mathcal{M}\}$. The market network is described as a graph $\mathscr{M}(M, E_M)$, where $E_M = \{\ (m, m') | m, m' \in M_s, \forall s \in \mathrm{N}\}$ is the set of connections between any two markets m and m' that co-occur in each start-up, and N represents the set of all start-ups in our database. The resulting network is undirected and weighted, the weight being equal to the number of times two markets appear in all the M_s subsets. The constructed network consists of 812 nodes and 18560 links.

In order to extract a hierarchical structure and define the macro categories of the start-ups' markets, I first "clean" the network by performing a backbone analysis [98]. This methodology consists of identifying the statistically relevant weighted links that must be preserved. Given a node i, each link shared by i with its neighbours is assigned with a normalized weight $p_{ij} = w_{ij}/s_i$, where w_{ij} represents the weight of the link between i and j, and $s_i = \sum_j w_{ij}$ represents i's strength. Each link is then associated with a probability $\alpha_{ij} = (1 - p_{ij})^{s_i - 1}$. The backbone network is populated by those links which satisfy $\alpha_{ij} < \alpha$, where α represents a confidence level that can be tuned in order to obtain stronger or weaker filtering.

A natural way to classify nodes in a network is through the use of network algorithms. The number of groups into which the nodes in the network will be divided is not defined *a priori*, but arises from the network structure. A community detection algorithm finds the most natural way to partition nodes into groups such that most of the connections between nodes fall within the same group, and only few across different ones. In this analysis, I chose to apply the Louvain Modularity community detection algorithm [99]. This algorithm optimizes the "network modularity", a real number ranging between -1 and 1, which measures the density of links of nodes

inside communities as compared to those outside. The optimisation of this value generally leads to the best categorisation of nodes in communities.

By applying the Louvain Modularity community detection algorithm to the unfiltered network, I obtain seven communities which show to be meaningless, i.e., market sectors that have nothing or very little in common appear to belong to the same category. For instance, *accounting*, *enterprise application*, and *call center* belong to the same community (or macro category).

When I apply the same algorithm to the extracted network's backbone, for $\alpha > 0.01$ the number of communities ranges from 7 to 20, while for $\alpha \leq 0.01$ the number of communities becomes more stable, with a value around 21 that does not significantly vary up to $\alpha = 10^{-5}$. The market tags that populate each of these 21 communities represent a meaningful node division. By increasing the statistical threshold from $\alpha = 0.01$ (i.e. decreasing α), the number of communities does not change while the number of edges and nodes in the network decreases. Therefore, in order to maximise the number of edges and nodes I choose as a statistical threshold $\alpha = 0.01$.

Of these 21 communities, nine of them present more than eight market tags. I associate with each of these nine communities a macro category which I manually label based on the market tag populating each community with the following names: *Curated Web, Education, Software for Enterprises, Data Analysis Software, E-commerce, Telecommunication, Finance, Advertising*, and *Leisure*. I incorporate all of the remaining smaller twelve communities into a broader category named "*Other*". A network representation with the different communities highlighted is shown in Fig.3.2, in which the colours represent the communities found by the Louvain Modularity detection algorithm and where nodes' sizes are proportional to the nodes' degrees.

3.3.2 Nation (and State) characterization based on markets

How can we make use of the market macro categories? I suggest to use them in order to characterize different nations based on the activity of start-ups located in them. In the following analysis I am considering a subset of nations with at least 2000 registered start-ups. This choice is driven by the fact that this subset of nations is

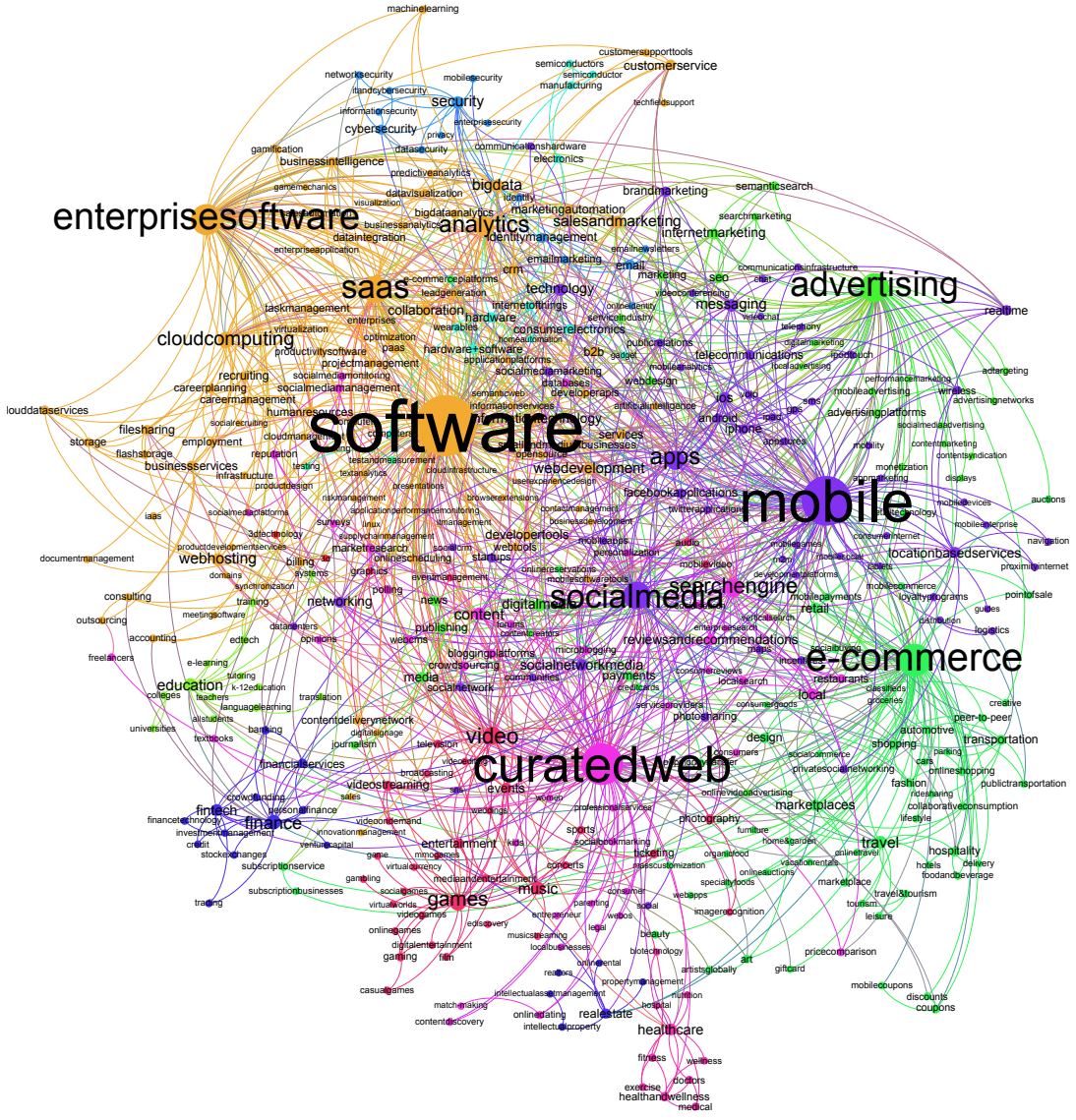

Figure 3.2: **The market network backbone.**

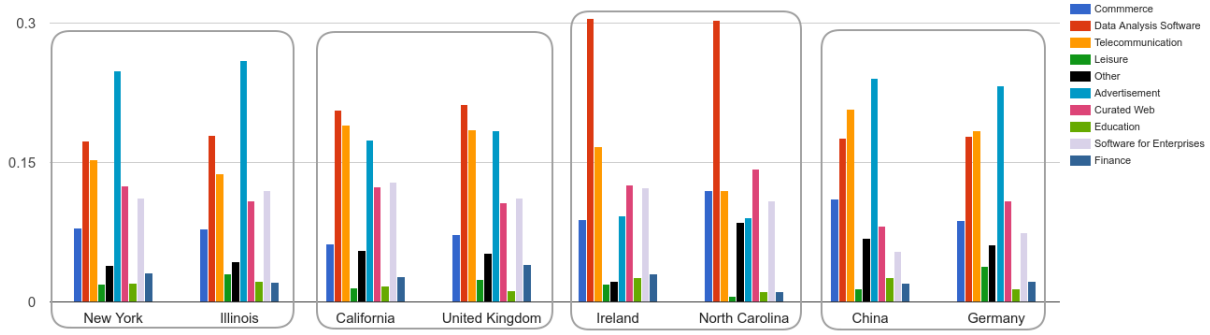

Figure 3.3: **Nations fingerprints based on markets**. Each histogram represents a nation fingerprint based on the market macro categories. In the figure are reported four pairs of nations/states which share the most similar fingerprints. The similarity is measured using the euclidean distance between all pairs of the 10-dimensional vectors associated with each nation.

sufficiently homogeneous in terms of number of start-ups and a comparison between them would result to be meaningful.

Given a national ecosystem, I assign each company's market to the corresponding market macro category (obtained from the community analysis on the backbone network). I then calculate and associate with each national ecosystem a vector containing the normalised distributions of macro market categories. An example of the normalised distributions for different nations is shown in Fig.3.3. This process enables a nation "fingerprint" to be identified, which reflects the unique pattern of the national start-up market's activity. In some nations, start-ups prefer to focus on the advertisement industry sector (e.g., the light-blue peak of New York and Illinois) while in other nations the major focus is on data analysis and software (e.g., Ireland and North Carolina). What is interesting to observe is that whilst nations show different fingerprints, geographically distant nations may show similar patterns. Fig.3.3 shows four pair of nations that share similar profiles.

Results suggest the presence of clusters of nations based on the similarities (and differences) of the market macro category distributions. Therefore, I compute the euclidean distance between the 10-dimensional vectors associated with each nation. To obtain an overview of all national ecosystem profiles, I produce a matrix in which each row represents the nation vector. By associating a color with each cell,

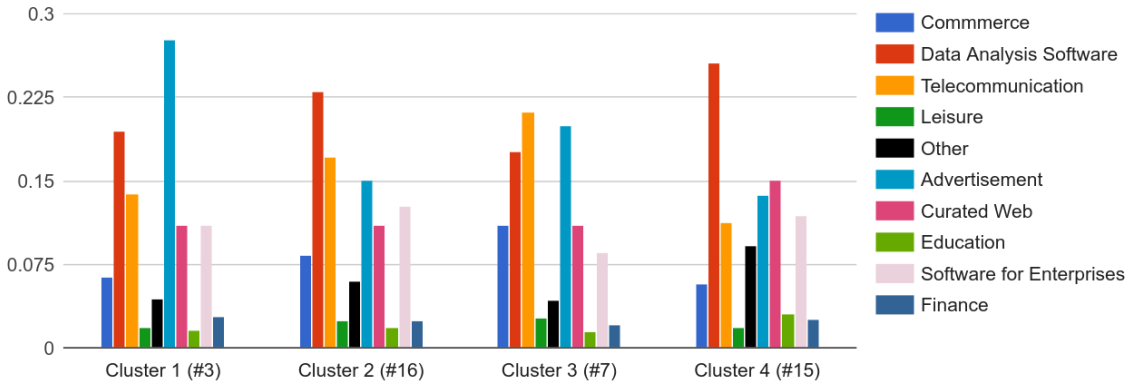

Figure 3.4: **Average cluster fingerprints.** Here are represented the average fingerprints constructed by producing the average distribution of the fingerprints of all nations that belong to the same cluster. The number in parenthesis (#X) indicates the number of nations within each cluster. Notice that each cluster presents a distinguished distribution that often are peaked around one of the market categories.

proportional to the cell value, I obtain the heat map shown in Fig.3.5. Storing the distance between all pairs of national vectors in a distance matrix, I perform a hierarchical clustering analysis. The clustering method used is complete linkage, and the associated dendrogram is shown on the left hand side of Fig.3.5. Results show four principal clusters.

The number of nation per cluster varies and it is reported next to the cluster number in Fig.3.4. The resulting averaged clusters show different distributions, with one or two leading market macro categories per cluster. For instance, cluster 1 (that includes New York and Illinois) is the less populated and shows a peak in its distribution corresponding to a preference for the *Advertisement* market sector. Cluster 2 (that includes Japan and Ireland) is the most populated cluster and has an homogeneous distribution among the markets categories with a preference for *Data Analysis Software* and *Telecommunication Services*. Cluster 3 (that includes Germany and China) is quite homogeneous with three leading markets, such as *Telecommunication*, *Advertisement*, and *Data Analysis Software*. Finally, Cluster 4 (that includes California and United Kingdom) is mostly focused in *Data Analysis Software*, followed by an equal interest in both *Curated Web* and *Advertisement*.

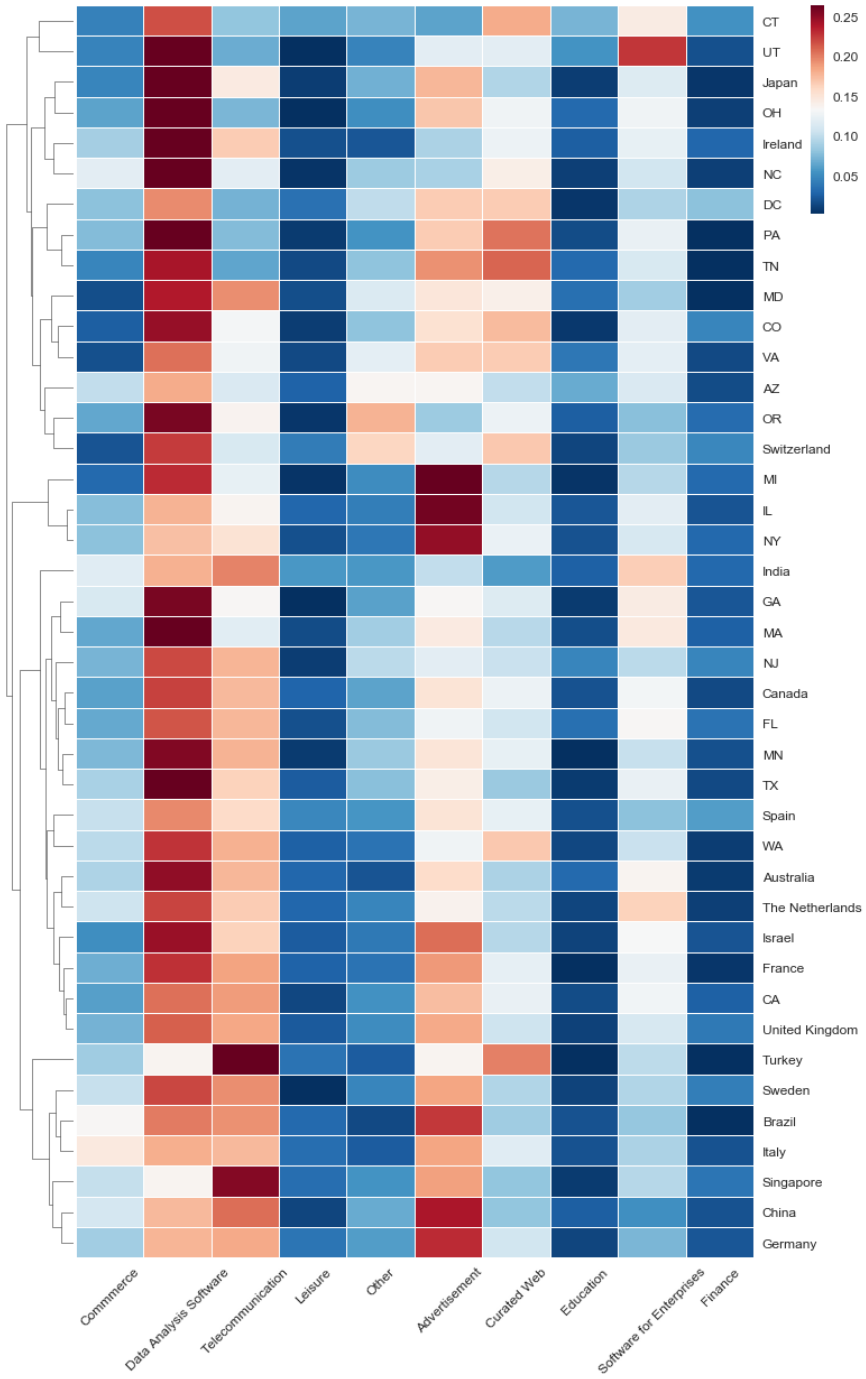

Figure 3.5: **Nations heatmap and dendrogram on markets clusters.**

3.4 Competition, mobility, and overlap networks

In this section I describe the construction of the two networks of start-ups and the resulting network made from their overlap. I then discuss the effect of competition on the mobility of employees among companies.

3.4.1 Declared-competition network

With the knowledge on the companies' competitors, it is possible to create a network in which a connection between a company i and a company j exists if the start-up i considers the start-up j to be its competitor. The resulting network is directed and it is possible to find that not all arcs are reciprocated. Such a network comprises of $N^{[1]} = 39,177$ companies and $E^{[1]} = 74,496$ arcs. I refer to this network as the declared-competition network. Data is aggregated over an observation period ranging from 1980 to 2016. Other information concerning the declared-competition network is reported in Table 3.2.

The total degree k_{tot} of a node is expressed as $k_{tot} = k_{in} + k_{out}$, being the sum of a node in- and out-degree respectively. The k_{tot} follows a power-law distribution $p(k_{tot}) \sim k_{tot}^{-\alpha}$ with an estimated value of the coefficient $\alpha = 3.04 \pm 0.08$. Fig.3.6 shows the k_{in} in-degree, k_{out} out-degree, and the k_{tot} total degree distributions with their corresponding α exponents. Notice that the values of α are in the typical ranges found in real world complex networks.

Moreover, I have measured and plotted the $k_{nn}^{tot}(k_{tot})$, i.e., the average total degree of the nearest neighbours of nodes with total degree k_{tot}, of the declared-competition network. In accordance with the results of Chapter 2, the trend is disassortative. In fact, the $k_{nn}^{tot}(k_{tot})$ negative trend is in agreement with what we would have expected for a network with competitive, i.e., negative relationships.

3.4.2 Mobility network

The second network I construct from CrunchBase data is the mobility network. In this network of start-ups, an arc from company i towards company j exists if at least one person leaving company i joins company j subsequently in his or her career (see left hand side of Fig.3.7). Such a network comprises of $N^{[2]} = 104,872$ companies and

Declared-competition network	
Nodes	$39,177$
Arcs	$74,496$
Density	5×10^{-5}
$\max(k_{total})$	580
$\max(k_{out})$, $\max(k_{in})$	365 342
Mobility network	
Nodes	$104,872$
Arcs	$158,824$
Density	1.4×10^{-5}
$\max(k_{total})$	1367
$\max(k_{out})$, $\max(k_{in})$	799 568
Overlap network	
Nodes	$16,781$
Arcs	280
$\max(k_{total})$	14
$\max(k_{out})$, $\max(k_{in})$	9 10

Table 3.2: **Networks statistics.**

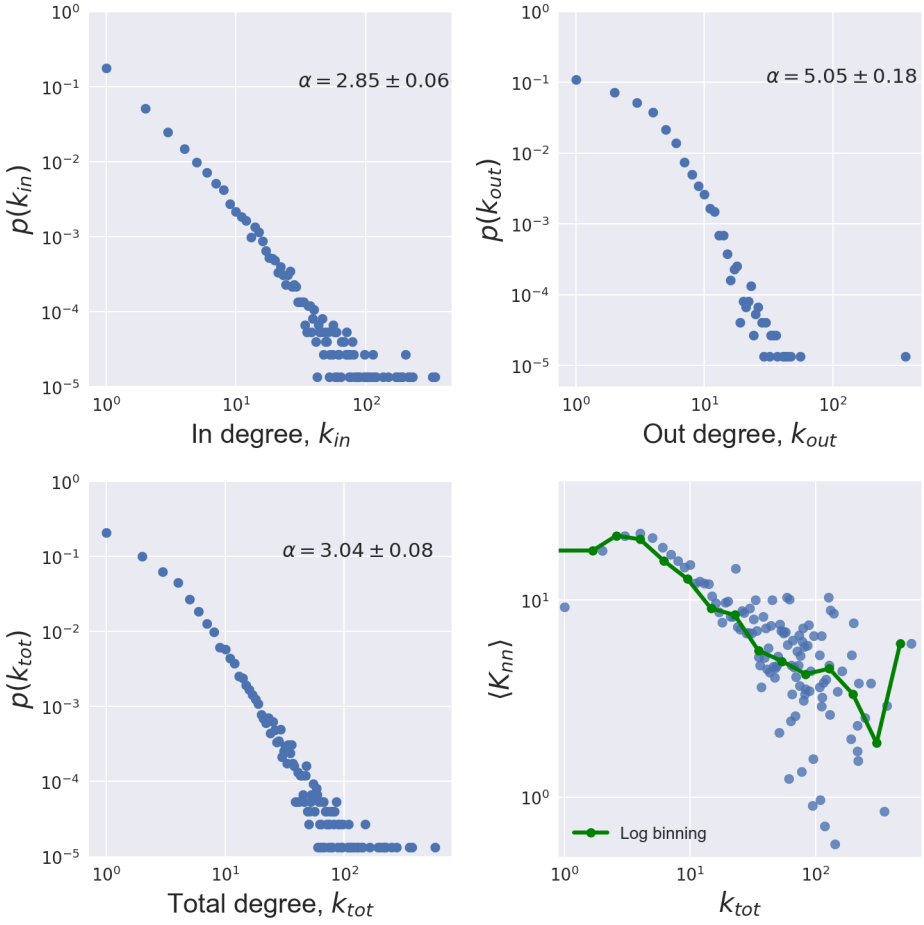

Figure 3.6: **The declared-competition network degree distributions and the disassortative trend.** In figure are reported the in-degree distribution $p(k_{in})$, the out-degree distribution $p(k_{out})$, the total-degree distribution $p(k_{tot})$ and the $k_{nn}^{tot}(k_{tot})$ distribution which gives an insight of the network disassortative trend. The log binned trend is shown in green.

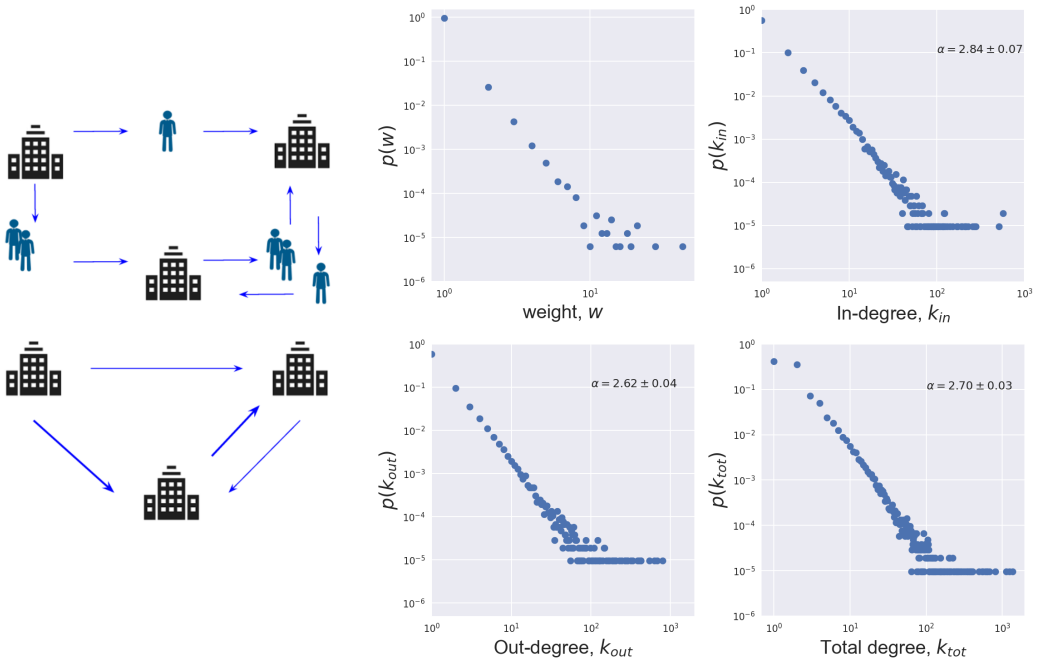

Figure 3.7: **The mobility network**.
Left hand side, top. Networks are created by considering the flow of employees between any two companies. Starting with a two-mode graph (top left figure) where companies are connected with their employees, I obtain the mobility network by producing the one-mode projection of this graph (bottom left figure).
Left hand side, bottom. A company is connected to another if there is the exchange of at least one employee. The resulting network is weighted, where the weight is equal to the number of employees that moved from a company to another.
Right hand side. Degree distributions and weight distributions for the mobility network.

$E^{[2]} = 158,824$ arcs. In this network arcs are weighted, where the weight is equal to the number of people that moved from one company towards another over time. This network is also characterized by a power-law total degree distribution $p(k_{tot}) \sim k_{tot}^{-\alpha}$ with an estimated value of the coefficient $\alpha = 2.70 \pm 0.03$ (see Fig.3.7 for all degree combinations and their exponent values). Other information concerning the mobility network can be found in Tab.3.2.

3.4.3 Overlap network

Do competitors exchange employees? In order to answer this question I produce the overlap between the previous two networks. I define the set of overlapping nodes as $N^{ov} = \{N^{[1]} \cap N^{[2]}\}$ (where $N^{[1]}$ is the set of nodes of the declared-competition network and $N^{[2]}$ is the set of nodes of the mobility one) and the set of arcs $E^{ov} = \{E^{[1]} \cap E^{[2]}\}$ (where $E^{[1]}$ is the set of arcs of the declared-competition network and $E^{[2]}$ is the set of arcs of the mobility one). If a link (i, j) from company i and company j exists in the mobility network but the link (j, i) does not and (j, i) exists in the declared-competition network but (i, j) does not, a connection between company i and j would not exist in the overlap network.

The resulting overlap network is constituted of $|N^{ov}| = 16,781$ nodes of which 76 belong to the weakly connected component with $|E^{ov}| = 107$ arcs. A representation of the weakly connected component is shown in Fig.3.8.

The tiny fraction of links should not be surprising. In fact, it is quite ordinary that in a contract of employment there are restrictions which prevent the employee from working for a competitor after leaving his current employment. This clause generally goes under the name of *restrictive covenant* or *restraint of trade*. Usually, a company is unable to hire an employee from one of its competitors for a limited period and within a limited area of work. However, big companies represented in Fig.3.8 (such as IBM, Oracle, Apple, Yahoo, and Google) do hire some employees from their direct competitors.

In order to extract some relevant information from this network, I produce the network measure defined as "attractiveness" [100]. De Domenico and Arenas have defined in [100] the measure of node attractiveness in a mobility network of scientists. This measure represents the ability of a nation (node) to attract people from its neighbours considering both the people that flow into and outside of the nation. Attractiveness is defined as

$$\mathcal{A}_i = \frac{s_i^{in} - s_i^{out}}{s_i^{out}}$$

where s^{in} and s^{out} represents respectively the in-going and out-going strength of the i-th node. The higher the attractiveness, the better a node is performing in respect of the others.

I reproduce the measure in the overlap network on companies with $s_{tot} = s_{in} +$

$s_{out} > 10$. In Fig.3.9 we observe that companies splits into two different blocks: those with positive attractiveness (blue), i.e. $s_{in} > s_{out}$ and those that have a negative one (red), i.e. $s_{in} < s_{out}$. Based on the actual dataset of companies, the inclusion of a temporal dimension in the study of attractiveness would have a negative impact on the network analysis. In fact, a temporal analysis would further filter the connections between companies, resulting in the creation of several sets of dyadic relationships. For simplicity and meaningfulness of the results, I prefer to report only a temporal aggregated analysis.

Google, HP, and Salesforce lead the ranking. It is reasonable to think that, based on the results, employees find in Google a good workplace compared to its competitors. Surprisingly, big companies like Apple, Yahoo, and IBM do not perform well in terms of attractiveness as their negative attractiveness value can be interpreted as the trend of employees in leaving the company in order to join a competitor.

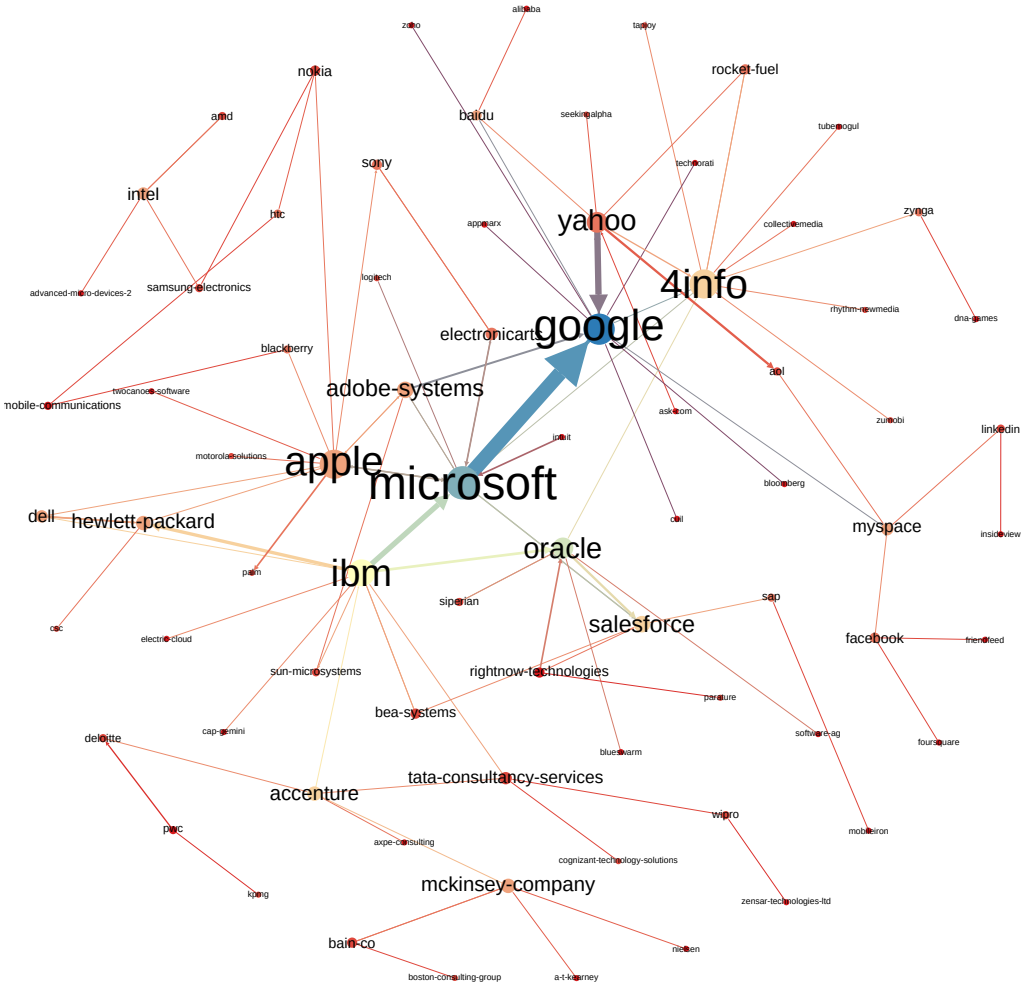

Figure 3.8: **Overlap between mobility and competition network**. Colours and size of a node are proportional to the nodes' k_{tot} (blue corresponds to a high k_{tot}, red to a low one). The arc's thickness is proportional to its strength.

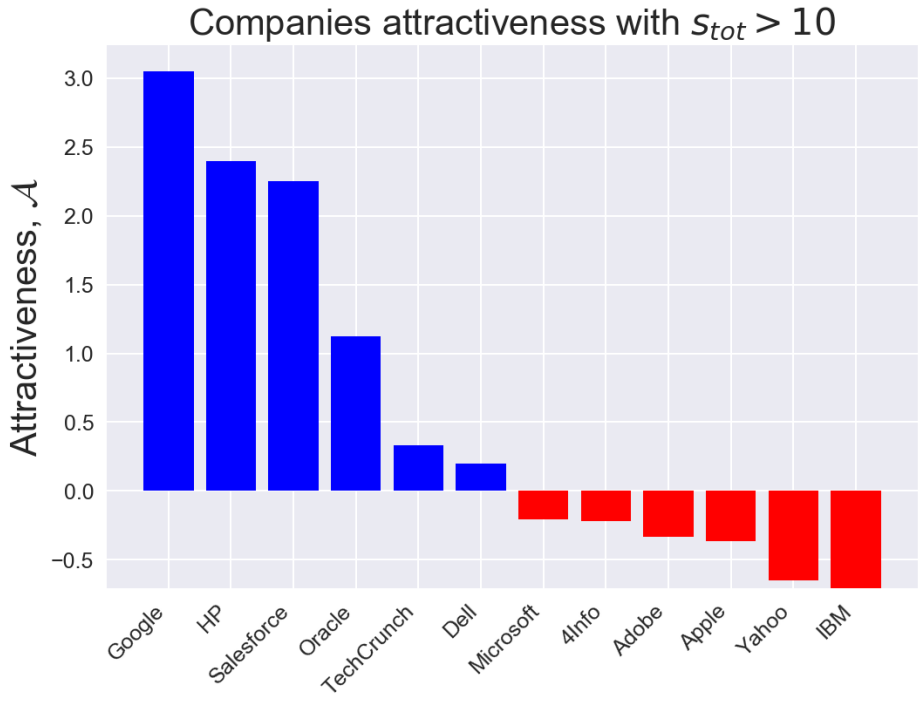

Figure 3.9: **Companies attractiveness**. Here are reported all companies belonging to the weak connected component of the overlap network for which $s_{tot} > 10$. Notice that companies splits into two different blocks: those with positive attractiveness (blue), i.e. $s_{in} > s_{out}$ and those that have a negative one (red), i.e. $s_{in} < s_{out}$.

3.5 Competition and success in national ecosystems

In this section, I move on to the analysis of national ecosystems, i.e., the set of start-ups whose headquarters are located within the same nation. First, I examine the flow of people into, out of, and inside each ecosystem. I then quantify the effects of competition on the mobility of people within an ecosystem, and I correlate the presence of a competitive environment with the success of the national ecosystems.

3.5.1 Flow of people among nations

For the following analysis I consider the set of nations and states (in the case of U.S.) in which at least 100 start-ups are located. The choice of considering U.S. states instead of the U.S. as a unique nation is driven by two main reasons: first, each state presents different economic regulations for start-ups; second, as shown in Section 3.3.2, each state can be considered as a different ecosystem with a specific pattern of start-up activity. Finally, a comparison between the U.S. and other nations based on quantities that are a function of the number of companies would give disproportionate (and less meaningful) results, being that the 55% of the registered companies in CrunchBase are based in the U.S..

I define with $\mathcal{N}_\nu$ the set of start-ups whose headquarters reside in a nation ν as $\mathcal{N}_\nu = \{i | i \in \nu\}$. Then, I construct ϕ, the weighted adjacency matrix of the mobility network among nations. This matrix can be seen as a coarse-grained matrix of the start-up mobility network. Each element represents the total flow of people that move from a company belonging to the set of start-ups $\mathcal{N}_\nu$ towards a company belonging to the set of start-ups $\mathcal{N}_\mu$ defined as

$$\phi_{\nu\mu} = \sum_{\substack{i \in \mathcal{N}_\nu, \\ j \in \mathcal{N}_\mu}} w_{ij}$$

where w_{ij} represents the weighted adjacency matrix element of the start-up mobility network. To express the flow of people that go from a nation ν towards other nations

such that $\nu \neq \mu$ I define

$$\phi_\nu^{out} = \sum_\mu \phi_{\nu\mu}.$$

Similarly, I define with ϕ_ν^{in} the in-going flow of people from other nations towards nation ν as

$$\phi_\nu^{in} = \sum_\nu \phi_{\nu\mu}.$$

Finally, I define with $\phi_{\nu\nu}$ the mobility of people within start-ups that belong to the same nation ν. Fig.3.10 shows the relations between the three flow quantities. In each figure the bisector is drawn (dashed line) in order to have an indicator of "equilibrium" between different types of flow (for instance, a point on the bisector in Fig.3.10 (a) has $\phi_\nu^{out} = \phi_\nu^{in}$). In Fig.3.10 (a) we observe an overall equilibrium between ϕ_ν^{out} and ϕ_ν^{in} with most of the national ecosystems lying on top the bisector. Fig.3.10 (b) shows a common trend for the majority of the ecosystems, with a predilection for most of the nations to have a $\phi_\nu^{out} > \phi_{\nu\nu}$. There are only a few exceptions, among which is the state of California (top right point). A similar result is shown in Fig. 3.10 (c) where, again, most nations have $\phi_\nu^{in} > \phi_{\nu\nu}$.

These results show that when people change their work, the change happens mostly across nations as opposed to within. This trend could be explained in view of what was previously explained concerning the restrictive covenant which disallow competitors from hiring an employee within a limited geographic area. However, it is not surprising that California, which is considered the cradle of start-up business with the Silicon Valley area, is more likely to offer people the possibility of changing start-up within nation (state) boundaries thanks to the presence of many and different start-ups, while in other nations the tendency is towards a change of location.

3.5.2 Competition and success in national ecosystems

In the introduction of this chapter I have defined the success of a start-up by assigning to it a binary success variable. Success is defined as a start-up having either undergone an IPO, acquired another company, or been acquired by another company. Here I extend the measure of success from a single start-up to the whole

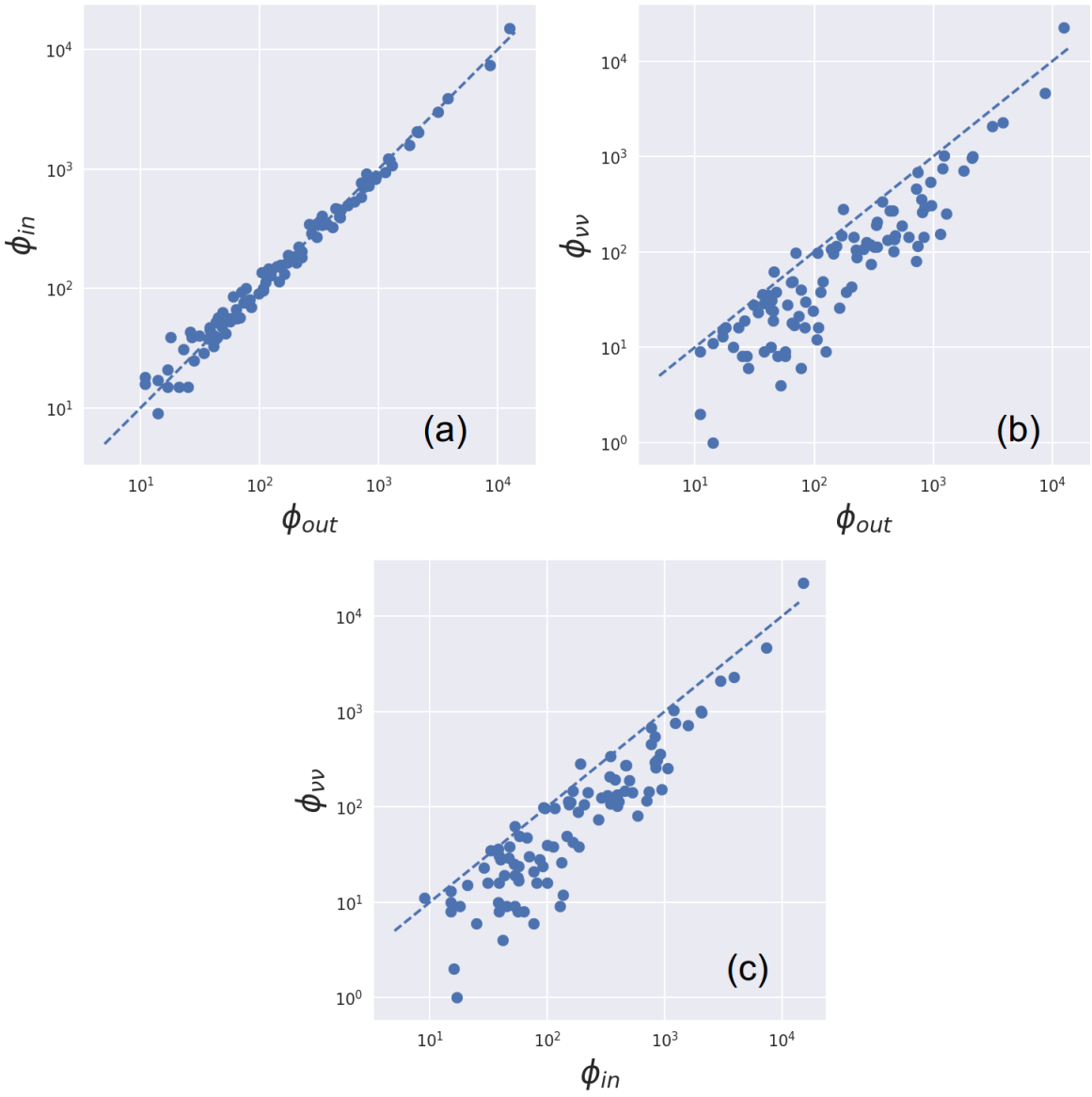

Figure 3.10: **Nations flow trends**.

national ecosystem.

With S_ν I denote the set of successful start-ups in a nation ν defined as

$$S_\nu = \{i \in \mathcal{N}_\nu \mid \quad i \text{ is successful}\}.$$

The most straightforward way to compare the success of different ecosystems would

be using the number of successful start-ups in the ecosystem. However, this approach would not allow us to understand which ecosystem is performing better in respect of the others because it does not take into consideration the size of the ecosystem. As a consequence, an ecosystem with only 10 start-ups of which 6 are successful would result performing less than an ecosystem with 10 successful start-ups out of 1,000. A natural way to evaluate the performance of an ecosystem considering its size could then be done by evaluating the ratio s_ν of the number of successful start-ups in respect of the total number of start-ups in that ecosystem $N_\nu = |\mathcal{N}_\nu|$, i.e.

$$s_\nu = |S_\nu|/N_\nu.$$

I want to quantify the extent to which the presence of competition between start-ups is related to the success of a national ecosystem. To this end, I define the quantity *competition blocking* $\mathcal{C}$ as the measure of how competitive a national ecosystem is as

$$\mathcal{C}(\nu) = \frac{\sum_{i,j \in \mathcal{S}_\nu} a_{ij}^{[1]}}{N_\nu(N_\nu - 1)}$$

where $a_{i,j}^{[1]}$ is the matrix element of the declared-competition network adjacency matrix, and the denominator represents the maximum number of connections that companies can make in a network with N_ν nodes. In network literature, this quantity is called network density. In this case, I assume that the denser the declared-competition network of a nation, the more competitive the ecosystem is.

I will now show that the previously defined quantity s_ν is not a suitable way to compare national ecosystems of different sizes. First, we need to look at the relationship between the number of successful start-ups S_ν and the number of companies N_ν for each nation ν. Fig. 3.11 (a) shows that the relationship between these two measures can be fitted by a power law as

$$Y(N_\nu) = Y_0 N_\nu^\beta$$

where β is the scaling exponent and Y_0 is a constant factor. The scaling exponent[5] $\beta = 1.26$ shows a superlinear trend, i.e., $\beta > 1$. A superlinear trend indicates

[5]The exponent is statistically relevant with a p-*value* $< 10^{-5}$

that the increase of a factor 10 in the number of start-ups in a nation ν is expected to result in an increase in the number of successful start-ups of a factor 10^{β}. For instance, in Fig.3.11 (a), the expected number of successful start-ups given a national ecosystem composed by $N_\nu = 1,000$ start-ups (which corresponds to the value $\log(1,000) = 3$ on the x-axis in figure) is approximately $|S_\nu| = 90$ (which corresponds to the value $\log(90) = 2.8$ on the y-axis in figure). Increasing the number of companies in an ecosystem by 10 times, i.e., $N_\nu = 10,000$, the expected number of successful companies is $|S_\nu| = 90 \cdot 10^{1.16} \simeq 1,300$, corresponding to the value 3.1 on the y-axis in figure. As a consequence, national ecosystems with N_ν companies that are shown to be above (below) the expected value $Y(N_\nu)$ are over (under) performing in respect of those whose value is $Y(N_\nu)$.

The national ecosystem's success rate s_ν does not properly take into account the non linear effects that are a consequence of the size of a national ecosystem. In fact, because of the presence of the superlinear trend, two national ecosystems that have both a success rate $s_\nu = 10\%$ may perform very differently in respect of one another, based on the number of start-ups N_ν. Moreover, we notice that in Fig.3.11 (a) many ecosystems have either a positive or a negative discrepancy between the expected value $Y(N_\nu)$ (solid line) and the value $|S_\nu|$ resulting from the data (dots). As suggested by Bettencourt *et al.* in [101], the suitable measure to compare entities of different sizes when dealing with superlinear scaling is the deviation from the expected value defined as

$$\xi_\nu = \log \frac{|S_\nu|}{Y_0 N_\nu^{\beta}}$$

which represents the performance of a nation ν compared to the one expected by the superlinear trend. Figure 3.11(b) shows the trend between ξ and $\mathcal{C}$. The two quantities follow a negative trend with a statistically validated exponent[6] $\beta = -0.19$. The trend suggests that a well defined correlation between the two quantities exists: the more competitive an ecosystem is, the less successful it is[7]. This result does not

[6]The exponent is statistically relevant with a p-*value* $= 6 \times 10^{-4}$

[7]As stated at the beginning of the section, considering as a measure of success only the number of successful start-ups in an ecosystem will give an incorrect representation of an ecosystem's performance due to the high heterogeneity of the ecosystems' population sizes. Concerning the quantity s_ν, i.e., the ratio between the number of successful start-ups and the total number of start-

imply causation, but opens the discussion to interpret the role of competition within an ecosystem.

ups in that ecosystem, I obtain a similar negative trend when correlating the measure with the competition blocking. Here I report the measure proposed in [101] as an ecosystem's performance measure because it is a suitable quantity for observations that follow superlinear trends.

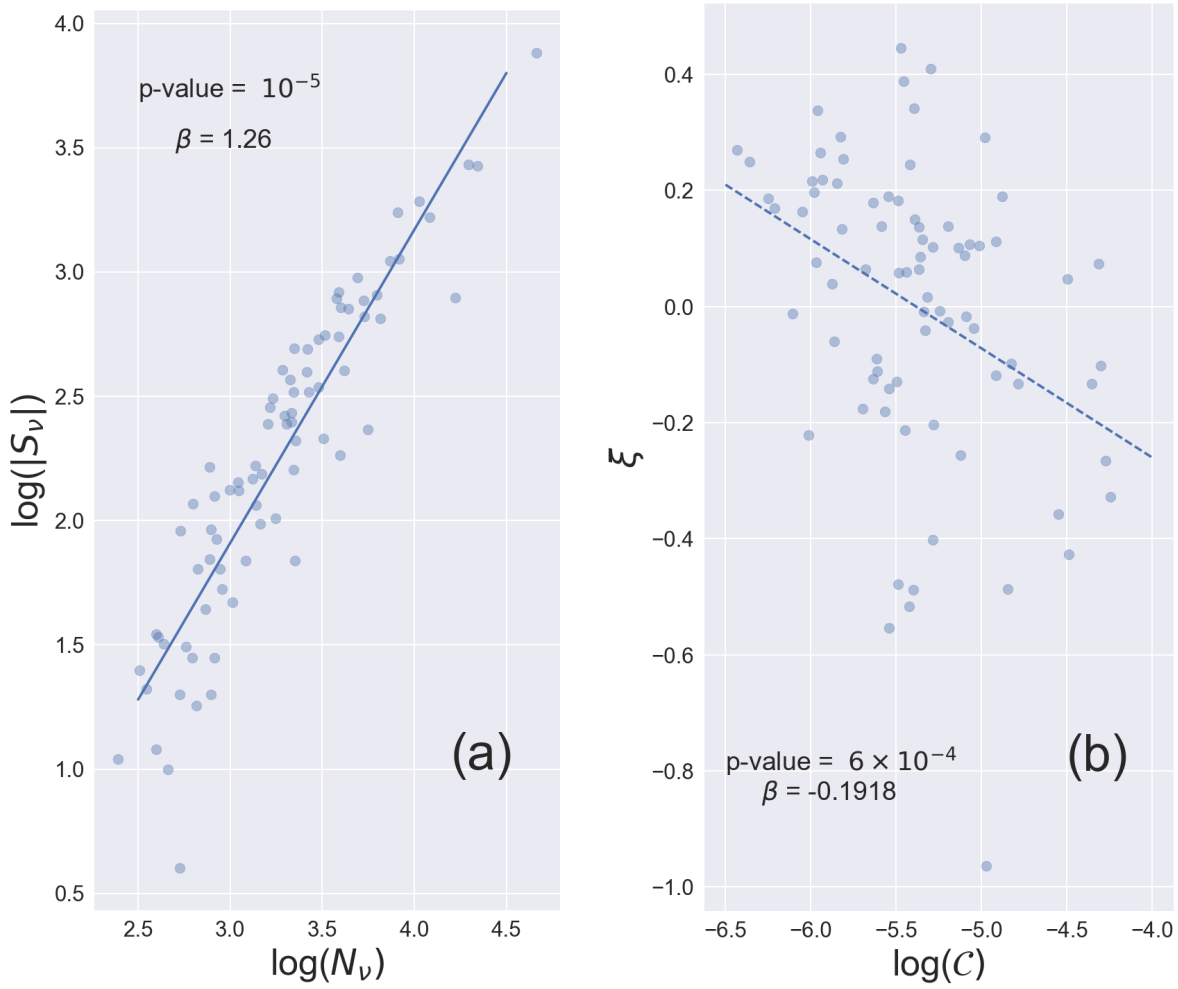

Figure 3.11: **Relations between number of start-ups, success, and competition blocking $\mathcal{C}$.**
(a) Superlinear scaling law (solid line) between the number of successful start-ups (blue dots) and the size of a national ecosystem. The slope of the solid line has a statistically relevant exponent $\beta = 1.26$.
(b) The negative trend between the residuals ξ versus the competition blocking factor. The slope of the dashed line has a statistically relevant exponent $\beta = -0.19$. Figure indicates that the higher the value of the competition blocking, the lower is the performance of the ecosystem.

3.6 Conclusions

In this chapter I have shown how, through the analysis of one of the most complete databases on start-ups, it is possible to extract information, describe, and characterise innovative ecosystems on a global scale. Start-ups represent a new and innovative way of running business which often brings new solutions, services, and products associated with markets that were not even imaginable a few decades ago (e.g. the apps market, the mobile market, etc...). Moreover, start-ups play an important role in creating job positions, with "*roughly two million to three million new jobs created every year*"[8]. Start-ups are distinguishable from larger companies by their investments in innovative technologies, while large companies prefer to put money into less risky incremental technologies. The start-ups' riskier strategy, when successful, brings to start-ups' investors a massive return on investment. However, to date, only one out of ten start-ups makes it through. It is therefore of key importance to understand what are the patterns that can help lead a start-up to success.

When it comes to characterising start-ups on a worldwide scale, a problem lies in the identification of the type of business or industrial sector in which they are involved. CrunchBase's dataset provides granular information (market tags) regarding the industrial sectors in which start-ups run their business. As these tags are produced by a free-text attribution method, they are often imprecise and possible of misspell. To make use of the information on the industry sectors provided by CrunchBase, I proposed a method to group market tags into macro categories. In order to do so, I have created the "weighted market network", a network in which market tags are nodes and a connection represents the co-occurrence of two market tags in at least one start-up. Finally, a weight is associated with each link equal to the number of times the market tags appears in a single start-up.

I then combined two network techniques in order to remove statistically irrelevant links (and nodes) and to identify the presence of network communities. By looking at the communities found by the algorithm, I manually labelled each community based on the industry sectors within it. These communities represent the market macro category. Then, I have characterised each nation through a vector that describes

[8] https://www.forbes.com/sites/petercohan/2011/06/27/why-start-ups-matter/#d4f50323620a

the distribution of markets associated with each start-ups that reside in it. Finally, I have analysed differences and similarities between nations and I have distinguished the various patterns of activity of different ecosystems, i.e. clusters of similar nations, through a hierarchical clustering method.

In the second part of this chapter I focused on the effect that competition among start-ups has on the mobility of employees between companies, and on the success of an innovation ecosystem. I was able to create and analyse the topological properties of start-up networks at a global level. I defined and created the declared-competition network among start-ups in which companies are connected if there is a competitive relationships among them, and the mobility network in which companies are connected with one another if there has been a flow of employees going from one company to the other. By performing the projection between these two networks I showed that the mobility of people between competitors is an exception, not a regularity. This could be due to the nature of employee contracts which, quite often, forbid employees from joining a direct competitor for a predetermined time.

Most importantly, I showed that the presence of competition between start-ups and the success of a national ecosystem are anti-correlated: ecosystems negatively perform in the presence of high competition within their national boundaries.

The approaches and methodologies developed in my research may be used to better understand how to help an innovative ecosystem to flourish. The free circulation of people between start-ups within an innovative ecosystem seems, when compared to the possible obstruction due to the presence of a high number of competitors, to play an important role in fostering success. In view of my results, nations and governments may be encouraged to sustain start-ups that prefer to collaborate, e.g. by removing some of the legal constraints against competitors in order to improve the success of these early stage companies. A flourishing and successful ecosystem may increase the capitalisation of start-ups, which could have a positive societal impact.

Chapter 4

Homophily and missing links in citation networks

In the previous chapter we saw that competition is a salient property of organization networks. However, competitive behaviour often emerges at the level of individuals. In this chapter (and in Chapter 5), I will focus on scientific research from the point of view of citations between scientific papers and collaboration among scholars. In particular, in this chapter I will analyse the patterns of knowledge flow among scientific articles based on the overlap of the articles bibliographies, focusing on the citations among papers.

Citation networks have been widely used to study the evolution of science through the analysis of knowledge flows among academic papers, authors, research sub-fields, and scientific journals. Furthermore, citations in science are also an important instrument to affirm the appreciation of a scholar's work. As a consequence, the omission of relevant citations may represent a way to undermine the prestige of a paper or even of an author. In this chapter I will propose a method that aims to uncover the absence of relevant citations but also the presence of irrelevant ones.

In order to define when the presence (or the absence) of a citation is relevant, I will analyse the citation networks of the American Physical Society (APS) journals dataset. First, I will test the presence of homophily (the social mechanism whereby the more similar two individuals are, the higher chance there is that they are connected) for knowledge transfer among papers. In order to achieve this, I will analyse

whether citations tend to occur between papers involved in similar topics or research problems. Then, I will propose a method for measuring the similarity between articles through the statistical validation of the overlap between their bibliographies. Results suggest that the probability of a citation made by one article to another is indeed an increasing function of the similarity between the two articles. This will enable missing citations between pairs of highly related articles, to be uncovered and may thus help identify barriers to effective knowledge flows.

By quantifying the proportion of relevant but missing citations, I will conduct a comparative assessment of distinct journals and research sub-fields in terms of their ability to facilitate or impede the dissemination of knowledge. Findings indicate that *Electromagnetism* and *Interdisciplinary Physics* are the two sub-fields in physics with the smallest percentage of missing citations. Moreover, knowledge transfer seems to be more effectively facilitated by journals of wide visibility and impact factor, such as *Physical Review Letters*, than by lower-impact ones. Hopefully, this study can have interesting implications for authors, editors and reviewers of scientific journals, as well as public preprint repositories, as it provides a procedure for recommending relevant but missing references and properly integrating bibliographies of papers.

4.1 Introduction

Among the broad category of information networks, including the Word Wide Web [14], email exchange networks [102], and phone call networks [103], the networks of citations between academic papers have been widely investigated to uncover patterns and dynamics of knowledge transfer, sharing, and creation in science [104–107]. The nodes of citation networks are academic papers, each containing a bibliography with references to previously published work. Typically, a directed link is established from one paper to another if the former cites the latter in its bibliography. Because papers can only cite other papers that have already been published, all directed links in citation networks necessarily point backward in time. Citation networks are therefore *directed acyclic graphs*, i.e., they do not contain any closed loops of directed links [108].

Since the seminal work by Derek de Solla Price on the distribution of citations

received by scientific articles [106,107], citation networks have extensively been studied to shed light on the mechanisms underpinning the evolution, diffusion, recombination, and sharing of knowledge over time [109, 110]. The reason why citation networks are crucial to understanding and modelling scientific production is clear. Although citations can serve different functions – for instance, they acknowledge the relevance of previous work, they help the reader of a paper to gather additional information about a specific topic, they point to related work or, sometimes, they can also express disagreement with, or level criticism against, a position endorsed in a paper [111] – the number of citations received is generally regarded as an indication of the relevance and quality of a paper as well as of its authors' prestige and scientific success [112]. Certainly, citation networks can be used to reconstruct the communication flows among different scientific communities and infer the relation among different research topics and sub-fields [112]. Recent work on citation networks has indeed proposed a new method for highlighting the role of citations as conduits of knowledge. For instance, Clough et al. [113,114] have proposed reduction methods to filter out the relevant citations preserving the causal structure of the underlying network of knowledge flows.

In this chapter, I study citations from a different perspective. Here I focus on citation networks to cast light on the salience of *homophily*, namely the principle that similarity breeds connection, for knowledge transfer between papers. First, I assess the extent to which the occurrence of a citation between two papers is driven by the similarity between them. Specifically, I investigate empirically a large data set of articles published in the journals of the American Physical Society (APS) [115], and I measure the similarity between any two articles by drawing on, and extending, a method originally proposed by Tumminello et al. in Ref. [116, 117] that enables to statistically validate the overlap between the bibliographies of the two articles. Results suggest that the number citations made by one article to another is indeed an increasing function of the similarity between the two articles. My findings thus indicate that the creation of links in citation networks can be seen as governed by homophily [43, 118–120].

Second, I propose a novel method for identifying missing links in citation networks. The gist of my argument is simple. I focus on pairs of articles characterised by high degrees of similarity; if a citation between them is missing, I regard the lack

of a directed link as a signature of a relevant yet unrecorded flow of knowledge in the network. By uncovering pairs of published articles with missing citations, I rank the APS journals and topics according to the incidence of missing data on knowledge flows.

This method has important implications for the analysis not only of published articles, but also of newly posted preprints on online archives, or of manuscripts submitted to scientific journals. Specifically, this method can be used to suggest interesting work and relevant literature that could, in principle, be included in the bibliography of recently posted or submitted preprints. As I witness a continuously increasing production of preprints and publication of new articles, it has become particularly difficult for authors to keep abreast of scientific developments and relevant works related to the domain of interest. As a result, lack of knowledge of prior or current related work and missing relevant citations may occur quite often. The method presented in this chapter can help the scientific community precisely to address this problem. In particular, it can be used not only by authors to integrate the bibliographies of their work, but also by editors of scientific journals to uncover missing citations and identify the appropriate reviewers for the papers they are considering for publication.

This chapter is organized as follows. In Section 4.2, I describe the data set In Section 4.3, I introduce and discuss the method for evaluating similarity between articles based on the statistical significance of the overlap between their respective bibliographies. In Section 4.4, I apply the proposed method to all articles published in the journals of the APS. I show that citations between articles are positively correlated with their similarity, and I then identify missing links between similar articles published in different fields and in different journals. In Section 4.5, I summarise the findings and discuss implications, limitations, and avenues for future work.

4.2 The APS data set

The APS data set includes bibliographic information on all the articles published by the American Physical Society between 1893 and 2009 [115]. The citation graph $G = (V, E)$ includes $|V| = 450,084$ articles, and $|E| = 4,710,547$ directed links. The citations refer only to articles that have been published on APS journals. For

each article I extracted the publication date, the main research subject (according to the PACS taxonomy), and its bibliography. Each article belongs to a specific journal. I restrict the analysis to the seven major journals, namely Physics Review A, B, C, D, E and Letter, which are specialised in different sub-fields of physics.

I performed the analysis at three levels, namely the entire citation network, the sub-graphs of the citation network induced by articles in each of the ten main sub-fields of physics, as identified by the highest levels of the PACS hierarchy, and the six sub-graphs induced by articles published in Physical Review Letters and in Physical Review A-E. In the analysis, I discarded articles that appeared in Review of Modern Physics, which publishes almost exclusively review articles. In Table 4.1 I report the description of the ten main categories in the PACS taxonomy and the topics covered by each of the six journals here considered.

4.3 Quantifying similarity between articles

Similarity between two articles can be measured in a number of ways. A straightforward, yet labour-intensive way of comparing articles is to semantically analyse their entire texts. Alternatively, similarity can be simply based on the co-occurrence of a few relevant concepts or keywords in the titles or abstracts of the articles. Moreover, similarity can be measured through the co-occurrence of classification codes, such as those included in the Physics and Astronomy Classification Scheme (PACS), which help identify the research areas to which each article belongs [121]. Here, I propose an alternative measure of similarity based on the comparison between the bibliographic lists of references included in two articles. The hypothesis is that, if two articles are concerned with related aspects of the same discipline or research problem, then their bibliographies will exhibit a substantial overlap. I shall therefore introduce a method for assessing the statistical significance of the overlap between the lists of references of two articles, and I shall then use the statistically validated overlap as as measure of the similarity between the two articles.

Table 4.1: The scientific domains associated with the PACS codes and journals

PACS code	**Domain**
00	General
10	The Physics of Elementary Particles and Fields
20	Nuclear Physics
30	Atomic and Molecular Physics
40	Electromagnetism, Optics, Acoustics, Heat Transfer, Classical Mechanics, and Fluid Dynamics
50	Physics of Gases, Plasmas, and Electric Discharges
60	Condensed Matter: Structural, Mechanical and Thermal Properties
70	Condensed Matter: Electronic Structure, Electrical, Magnetic, and Optical Properties
80	Interdisciplinary Physics and Related Areas of Science and Technology
90	Geophysics, Astronomy, and Astrophysics
Journal	**Domain**
Physics Review A	Atomic, molecular, and optical physics
Physics Review B	Condensed matter and materials physics
Physics Review C	Nuclear physics
Physics Review D	Particles, fields, gravitation, and cosmology
Physics Review E	Statistical, non-linear, and soft matter physics
Physics Review Letter	Moving physics forward

4.3.1 Overlap between reference lists as a measure of similarity between articles

A natural way to quantify the overlap between two given sets Q_i and Q_j is the Jaccard index, which is defined as the ratio between the cardinality of the intersection in the two sets and the total number of elements in the union of the two sets:

$$J_{ij} = \frac{|Q_i \cap Q_j|}{|Q_i \cup Q_j|}. \tag{4.1}$$

Notice that, in general, if two sets share a higher number of elements, then their Jaccard index will increase, and in particular $J_{ij} = 1$ only if $Q_i \equiv Q_j$, while $J_{ij} = 0$ if the two sets do not share any element. An example of the suitability of the Jaccard index for measuring the similarity between the bibliographies of two articles is provided in Fig. 4.1(a)-(b). Here the two sets Q_i and Q_j represent, respectively, the articles in the two reference lists of the two articles i and j. Since article P1 and article P2 share only one reference over a total of five, their Jaccard index is equal to 0.2. Conversely, the two articles P3 and P4 in panel (b) have a Jaccard index equal to 1.0, since the overlap between their reference lists is complete.

However, the use of the Jaccard index has some drawbacks. First, the value of J_{ij} is always bounded from above by $\frac{\min(|Q_i|,|Q_j|)}{|Q_i|+|Q_j|}$. This means that if the sizes of the two sets are remarkably different, their similarity is primarily determined by the size of the smallest of the two sets. As a consequence, large sets tend to be characterised by relatively small values of similarities with other smaller sets. In addition to this, the Jaccard index does not distinguish between pairs of identical sets having different sizes. In particular, if I consider two identical sets (Q_i, Q_j) of size N_1 and two other identical sets (Q_m, Q_n) of size N_2, then I have $J_{ij} = J_{mn} = 1$, regardless of the values of their sizes N_1 and N_2. For instance, the Jaccard index of articles P5 and P6 is equal to 1.0 and is identical to that of articles P3 and P4, even though P3 and P4 share a larger number of references. In the case of bibliographic references, this degeneracy of the Jaccard index is very important. In fact, if I interpret references as proxies for knowledge flows from cited to citing articles, then it would be reasonable to associate a higher value of similarity to a pair of articles that share a large number of references than to a pair sharing only few references,

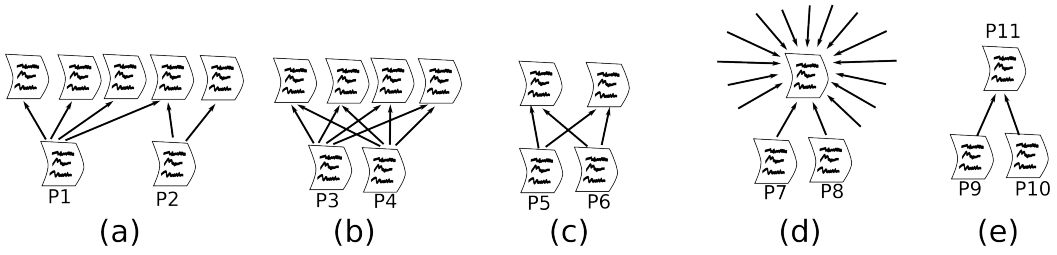

Figure 4.1: **Quantifying the similarity between two articles based on their bibliographies.** The similarity between two articles can be defined in terms of the overlap between their reference lists. The two articles P1 and P2 in panel (a) share only one citation; they should therefore be considered less similar than articles P3 and P4 in panel (b) which share four citations. This difference can be captured by the Jaccard index, which is equal to 0.2 in the former case and to 1.0 in the latter. However, the Jaccard index is equal to 1.0 also for the two articles in panel (c), which instead share only two citations. If citations are interpreted as proxies for knowledge flows, then the similarity between article P7 and P8 in panel (d), which cite a highly-cited article, should be smaller than the similarity between articles P9 and P10 in panel (e), which instead are the only two articles citing P11. The similarity measure, based on statistical validation, properly takes these heterogeneities into account.

since the former pair is expected to draw on a more similar scientific background. In particular, I would expect the two articles in panel (b) to be assigned a value of similarity larger than the two articles in panel (c).

Another drawback of a bare count of the number of common references is that some citations can, in principle, be more important than others. Consider the two cases depicted in Fig. 4.1(d)-(e). In panel (d), articles P7 and P8 have an identical set of references, consisting in the citation to a single highly-cited article. Also in panel (e), both articles P9 and P10 cite the same article. However, in this case the cited article does not receive any citation from other articles. Now, since the aim is to quantify the similarity between articles, a citation to a highly-cited article, such as a review article, should be considered less relevant than a citation to a more specialised or less visible article, which is cited only by articles concerned with a certain specific topic. In other words, it would be preferable to associate a higher relevance to the single citation shared by articles P9 and P10 in Fig. 4.1(e) than to the citation to other highly cited articles shared by articles P7 and P8 in Fig. 4.1(d),

and thus to conclude that articles P9 and P10 are more similar than article P7 and P8.

4.3.2 Defining statistically significant bibliographic overlaps

The method I propose here allows to overcome the drawbacks of the Jaccard index discussed above and illustrated in Fig. 4.1. The method is based on an extension of the so-called *Statistically Validated Network (SVN)* approach to the case of directed unipartite graphs. Statistically Validated Networks were introduced by Tumminello et al. [116, 117] as a method to filter out statistically irrelevant information from bipartite graphs, such as user-item networks deriving from purchase systems or product reviews. In such systems, a set A of nodes (e.g., buyers, users) express preferences over another set B of nodes (e.g., books, movies, services). Those preferences or selections are represented by directed links from nodes in set A to nodes in set B. The idea behind SVNs is that the similarity between two nodes i and j in the set A can be expressed in terms of the co-occurrence of their selections of nodes in B, and in particular that it is possible to attach a statistical significance, namely a p-value, to each set of common selections made by i and j.

Citation networks are not bipartite graphs. They are also different from user-item networks because each article in general can only cite other articles that have already been published, and can only receive citations from other articles that will be published after its publication date. Nevertheless, it is possible to draw upon the same idea used to construct bipartite statistically validated networks, and define a similarity between two articles based on the overlap between their reference lists.

Let's consider two sets of nodes, A and B. The set A contains all the articles with more than zero outgoing citations, $A = \{i \in V \,|\, k_i^{\text{out}} > 0\}$, while the set B contains all the articles that have received at least two citations, $B = \{i \in V \,|\, k_i^{\text{in}} > 1\}$. It is worth noticing that $A \cap B \neq \emptyset$, i.e., the two sets may share some articles, since in general each article cites and is cited by other articles. I denote by $N_A = |A|$ and $N_B = |B|$ the cardinality of the two sets. The method associates a statistical significance to the similarity between a pair of nodes (i, j) in A by comparing the number of co-occurrences of citations in their reference lists against the null hypothesis of random co-occurrence of citations to one or more articles in

B. In this way, the method allows to identify pairs of nodes in A characterised by overlaps between citations to elements in B which are statistically different from those expected in the null model.

The method works as follows. For each value k of in-degree observed in the citation network, I consider the set of nodes $S^k = S^k_B \cup S^k_A$, where $S^k_B \subset B$ contains all $N^k_B = |S^k_B|$ articles with in-degree equal to k, and $S^k_A \subset A$ contains all articles that cite at least one element in S^k_B. Notice that the set S^k is, by construction, homogeneous with respect to the in-degree of the elements belonging to the set B. Then, for each pair of articles $i, j \in S^k_A$, I indicate by d_i and d_j their respective number of citations directed towards the elements of S^k_B. Under the hypothesis that the articles i and j cite, respectively, d_i and d_j distinct elements uniformly at random from S^k_B, the probability that they select the same X articles is given by the hypergeometricprobability function:

$$\mathcal{P}(X \mid N^k_B, d_i, d_j) = \frac{\binom{d_i}{X}\binom{N^k_B - d_i}{d_j - X}}{\binom{N^k_B}{d_j}}. \tag{4.2}$$

Thus, I can associate a p-value to each pair of nodes $i, j \in S^k_A$:

$$q_{ij}(k) = 1 - \sum_{X=0}^{N^k_{ij}-1} \mathcal{P}(X \mid N^k_B, d_i, d_j), \tag{4.3}$$

where N^k_{ij} is the measured number of references that i and j have in common in the set S^k_B. The p-value, $q_{ij}(k)$, is therefore the probability that the number of articles in the set S^k_B that both i and j happen to jointly cite by chance is N^k_{ij} or more. I repeat the procedure for all possible values of in-degree k from $k_{\min}$ to $k_{\max}$, so that each pair of articles (i, j) is, in general, associated with several p-values, one for each value of in-degree k of the articles in their reference lists. Once all the p-values have been computed, I set a significance threshold p^* and validate all the pairs of nodes that are associated with a p-value smaller than the threshold p^*. Given a value of the statistical threshold, only the validated pairs of articles are considered similar at that significance level.

However, because each pair of articles (i, j) can be associated with multiple p-values, it is necessary to perform hypothesis-testing multiple times. In this case,

if I choose a confidence level or significance threshold p^*, say 1% confidence level ($p^* = 0.01$), the various p-values associated with the same pair of nodes are not compared directly with the chosen significance threshold p^*, but with a rescaled threshold that appropriately takes the number of tests performed into account. As a method for multiple testing I used the False Discovery Rate (FDR) [116, 122] (see Appendix for details).

The use of the hypergeometric probability function can be used to evaluate statistically significant citations in a citation network due to the heterogeneity of the number of received citations k which follow a power-law distribution. This methodology could fail for networks in which links are homogeneously distributed such as random networks.

Ultimately, I identify the set $\mathcal{M}(p^*)$ of all pairs of nodes whose similarity is statistically significant at the confidence threshold p^*. In what follows, I shall denote by $M(p^*) = |\mathcal{M}(p^*)|$ the cardinality of such set. In principle, since each pair of articles (i, j) can belong to different sets S^k (and, as a result, can be associated with several p-values $q_{ij}(k)$), it would be possible to define a similarity weight $w_{ij}(p^*)$ for each pair (i, j) as the number of times that the pair is validated at the confidence threshold p^*. In other words, $w_{ij}(p^*)$ would be the number of sets S^k for which $q_{ij}(k)$ passes the statistical test. However, I do not consider this possibility here, but simply assume that a pair of articles (i, j) belongs to the set $\mathcal{M}(p^*)$ if at least one of the p-values $q_{ij}(k)$ passes the statistical test at the confidence threshold p^*.

Notice that the definition of the p-value associated with a pair of articles in terms of the hypergeometric null model provided in Eq. 4.2 does not depend on the order in which two articles are assessed. The resulting symmetric value of similarity between any two articles is rooted in the invariance of the hypergeometric distribution in Eq. 4.2 under permutation of the pair i and j, i.e., of the two quantities d_i, d_j. Moreover, Eq. 4.2 rectifies some of the problems of measures of similarity based on a bare count of co-occurrences. In particular, two articles that share a small number N_{ij}^k of citations will be assigned a higher p-value (i.e., a smaller statistical significance of their similarity) than two articles sharing a large number of citations. This means that, for instance, the p-value $q_{P3,P4}(2)$ associated with the pairs of articles $(P3, P4)$ in Fig. 4.1(b) will be smaller than the p-value $q_{P5,P6}(2)$ associated with the pair of articles $(P5, P6)$ in Fig. 4.1(c), since $P3$ and $P4$ share a larger

number of references (namely, four instead of two) to other articles each receiving two citations. Moreover, the p-value associated with the pair $(P7, P8)$ will be larger (i.e., the similarity between the pair is less statistically significant) than the p-value associated with the pair $(P9, P10)$. The reason lies in the fact that, according to the hypergeometric null-model, the co-occurrence of a reference to a highly-cited article is more likely to take place by chance than the co-occurrence of a reference to an article with a relatively small number of citations.

4.4 Results

I now show how the proposed method for assigning a statistical significance level to the similarity between any pair of articles based on the statistically validated overlap between the respective bibliographies can indeed turn very useful and help uncover important properties of a citation network.

As an example of the possible applications of the method, I analyse the citation network among articles published in the journals of the APS during the period between 1893 and 2009. The data set is described in detail in Section 4.2. I shall start by studying empirically the probability $P_{i\to j}(p^*)$ of the occurrence of a citation from an article i to an article j validated at a certain statistical threshold p^*. I shall then discuss how the method can be used to identify missing and potentially relevant references and also to rank journals and scientific topics based on the relative occurrence of missing citations.

4.4.1 Homophily in citation patterns

I start from the observation that if I consider progressively smaller values of the statistical threshold p^*, the set $\mathcal{M}(p^*)$ will shrink and contain only pairs of articles characterised by an overlap between bibliographies that is highly significant, since it has passed a more stringent statistical test. Thus, small values of p^* single out pairs of articles that have a highly significant combination of common cited articles. But if two articles share significantly similar bibliographies, then there is a high probability that they are concerned with the same topic or research problem. As a result, it would be reasonable to expect a citation to occur from the more recently

published article to the one published at an earlier date. For each value of the statistical threshold p^*, I computed the number of pairs of articles $M(p^*)$ validated at that threshold in the APS citation network, and the number $K(p^*)$ of existing citations between those validated pairs. Then, I define the probability $P_{i\to j}(p^*)$ that there exists a citation between any two articles whose similarity is validated at the threshold p^* as:

$$P_{i\to j}(p^*) = \frac{K(p^*)}{M(p^*)}. \tag{4.4}$$

The obtained values of $P_{i\to j}(p^*)$ are reported in Fig. 4.2 as a function of p^*. The plot clearly suggests that the probability of finding a citation between two articles characterised by a highly statistically significant overlap between the respective reference lists (i.e., the similarity between that pair of articles is validated at a small value of p^*) is higher than the probability of finding a citation between articles whose reference lists are only moderately significantly similar. For instance, a citation between a pair of articles (i, j) whose overlap between reference lists is validated at $p^* = 10^{-2}$ occurs only with probability $P_{i\to j} \simeq 0.35$, while citations occur within up to 73% of the pairs of articles validated at $p^* = 10^{-7}$. In other words, the probability that an article i cites another article j is an increasing function of the similarity between the two articles.

In the social sciences, the principle that similarity breeds connection is traditionally referred to as homophily. This principle has been documented in a variety of empirical domains [43,118–120]. It is interesting to observe that homophily can also be found to govern citation networks where it plays an important role in shaping the structure and evolution of knowledge transfer between academic papers.

4.4.2 Suggesting missing references

The identification of a statistically significant similarity between two articles can be used to uncover potentially missing references. For instance, the implementation of a recommendation procedure based on statistically significant overlaps between bibliographies might be useful to assist the editor of a scientific journal in suggesting a list of possibly relevant (and missing) references to the authors of a submitted paper.

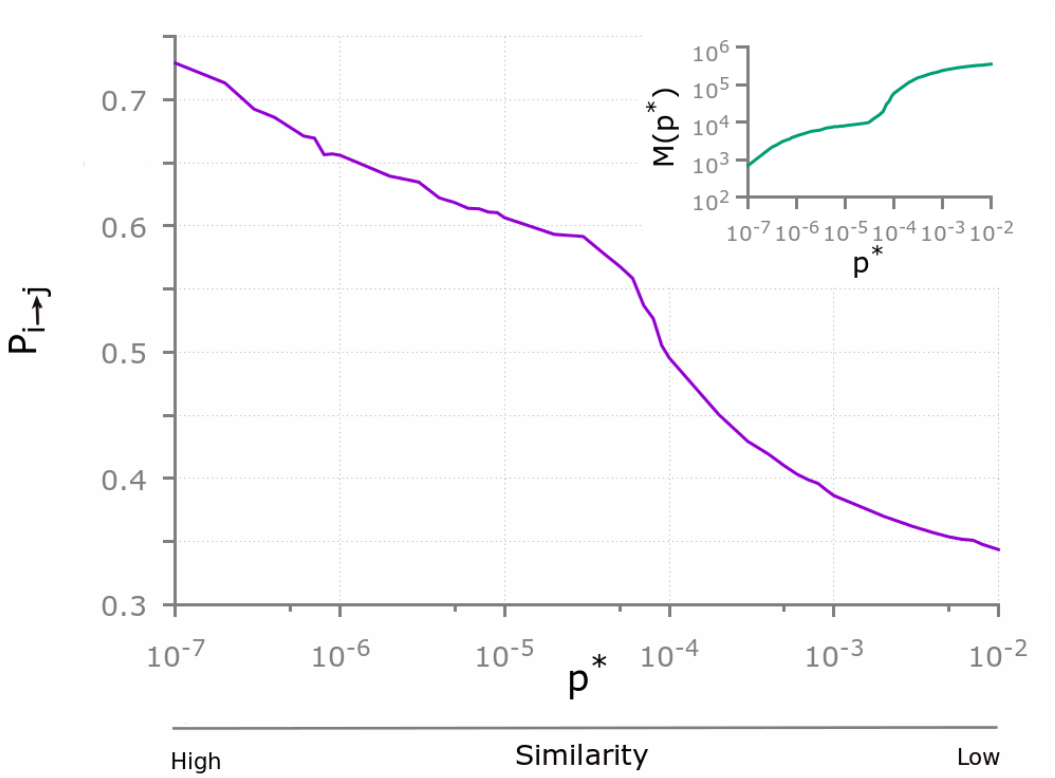

Figure 4.2: **The probability $P_{i\to j}(p^*)$ to observe a citation between two articles whose bibliographies overlap is statistically significant at the threshold value p^*.** Notice that $P_{i\to j}(p^*)$ increases as the statistical threshold p^* decreases. That is, citations between pairs of articles characterised by a highly significant overlap tend to occur with a higher likelihood than citations between articles whose reference lists are not significantly similar. The inset shows how the number of pairs of articles characterised by a statistically significant similarity at a given threshold p^* varies with p^*.

Fig. 4.3 shows a typical problem that could be fruitfully addressed through an appropriate reference recommendation system based on the identification of statistically significant overlaps between bibliographies of papers. I report a subgraph of the APS citation network consisting of several pairs of articles validated at $p^* = 10^{-7}$, the highest statistical. Each article is represented as a node, and validated pairs of nodes are connected through a link. The color of each link indicates whether the older article was (green) or was not (red) cited by the more recent one. Note that there is a prevalence of green links, which is consistent with the fact that, for a significance level $p^* = 10^{-7}$, a citation between a validated pair of articles occurs in more than 73% of the cases (see Fig. 4.2). However, I notice that article A has a considerable number of missing citations, resulting from the fact that it was not cited by any of the four articles that were published after its publication date and with which it shares a statistically significant portion of its bibliography (namely, nodes C, D, E, F). This could mean that either the authors of articles C-F were not aware of the existence of article A, despite the substantial overlap between their reference lists, or that article A was not particularly relevant to the topics addressed in the other articles.

Surprisingly, a more in-depth analysis of the articles in Fig. 4.3 suggests that, not only did all of them appear in the same journal (Physical Review E), but indeed they are all concerned with the same topic (electric discharges) and share a relatively large fraction of PACS codes (05.45.-a, 52.80.Hc). The high degree of similarity between topics can also be easily inferred from the abstracts and introductions of these articles. Interestingly, I found that articles B-F (yellow nodes) were all co-authored by the same research group G_1, while article A (the only blue node) was the result of the work of a different research group G_2. The fact that also article A does not cite article B suggests that the researchers in group G_1 were likely to be unaware of the work conducted by group G_2 in the same research field, and vice-versa.

In this particular case, the quantification of statistically significant overlaps between bibliographies could have been used to facilitate the flow of knowledge between different research groups. For instance, the editor of Physical Review E or the selected reviewers could have brought article B to the attention of the authors of article A, and similarly, when articles C-F were submitted to the same journal, the editor or the reviewers could have advised the authors of group G_2 to include article

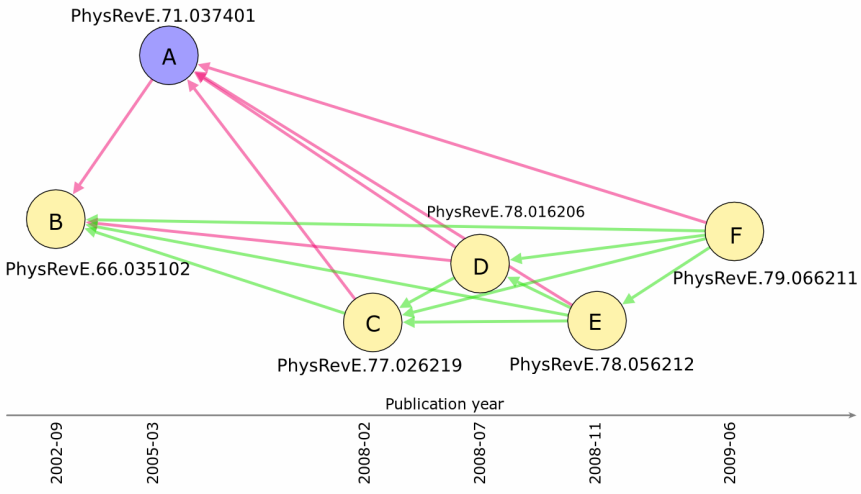

Figure 4.3: **Lack of knowledge flows.** An example of several validated pairs of articles in the APS citation network at $p^* = 10^{-7}$ (articles are reported in increasing order of publication time, from left to right). The occurrence of a link indicates that the pair of articles has passed the statistical test, while the colour of the link indicates that the most recent article in the pair actually did (green) or did not (red) cite the other one. In this case, all the articles represented as yellow nodes are articles co-authored by researchers in the same group, while article A was co-authored by another group. The identification of a large number of missing citations suggests that the two groups might have been unaware of the work of their colleagues in the same field.

A in the bibliographies of their submitted papers.

In the example reported, I have chosen to analyse the presence (and absence) of highly similar papers. The choice of filtering papers in respect to a single (specific) value could be relaxed by using binned ranges of statistical thresholds which could be defined, for example, as "high" similarity (e.g. $p^* \in (0, 10^{-6}]$), "medium high" similarity (e.g. $p^* \in (10^{-6}, 10^{-5}]$), and "medium low" similarity (e.g. e.g. $p^* \in (10^{-5}, 10^{-4}]$). This qualitative association could possibly help in an automation of the identification of relevant missing citations.

4.4.3 Ranking journals and disciplines by (lack of) knowledge flows

So far the analysis has been focused on the whole APS citation network. Physics is a very broad disciplinary area, including sub-fields as diverse as atomic physics, astronomy, particle physics, statistical mechanics, just to mention a few [112]. It is therefore reasonable to perform the analysis of the probability $P_{i\to j}(p^*)$ at the level of sub-fields. Specifically, I argue that the percentage $P_{i\to j}(p^*)$ of citations occurring between pairs of articles associated with a similarity that is validated at the statistical threshold p^* can serve as a proxy for the knowledge flows taking place within a sub-field. In what follows I restrict the analysis to the six citation sub-graphs induced by the articles published in each of the six research journals published by APS (in order to quantify the ability of each journal to facilitate or impede the dissemination of knowledge), and to the ten sub-graphs associated with the highest levels in the PACS taxonomy (which could shed light on the typical patterns of knowledge dissemination in different sub-fields). The lack of knowledge flows within a journal or a sub-field at a certain confidence level p^* can be quantified by the fraction of missing links:

$$U(p^*) = 1 - \frac{K(p^*)}{M(p^*)} = 1 - P_{i\to j}(p^*). \tag{4.5}$$

In general, the lower the value of $U(p^*)$, the more likely it is that a citation occurs between a pair of articles characterised by a similarity validated at the statistical threshold p^*. Fig. 4.4(a)-(b) shows how $U(p^*)$ behaves as a function of p^*,

respectively, for all articles whose main PACS code is either in group 40 (Electromagnetism) or in group 50 (Gases and Plasmas), and for all the articles published in Physical Review Letters and in Physical Review C. The figure clearly shows that, even though in all cases $U(p^*)$ decreases when $p^* \to 0$, different journals and different sub-fields tend to be characterised by slightly different profiles of $U(p^*)$, namely by different propensities to obstruct knowledge flows between similar academic papers. A comparative assessment of journals and sub-fields according to their typical ability to facilitate the dissemination of knowledge would, of course, be based on $\frac{K(p^*)}{M(p^*)}$. Moreover, the ranking will in general depend on the chosen value of the statistical threshold p^*.

From a theoretical point of view, a suitable approach to the ranking would be to compute the quantity:

$$U_0 = \lim_{p^* \to 0} U(p^*), \tag{4.6}$$

namely the limiting value of $U(p^*)$ when I let the statistical threshold p^* go to zero. However, this quantity cannot be computed accurately for a finite network, since for a certain value $p^* > 0$ the number $M(p^*)$ of validated pairs at p^* will be equal to 0, and the ratio $\frac{K(p^*)}{M(p^*)}$ would therefore be undetermined. Here I employ a simple workaround, namely I consider the tangent at the curve $U(p^*)$ at the smallest value of p^* for which the number of validated pairs is still large enough for the construction of a network of a reasonable size (I found that 10^{-7} is an appropriate choice in this case), and I compute the intercept at which this tangent crosses the vertical axis. This method could fail in two cases (which were not encountered in this work): i) if at $p^* = 10^{-7}$ there are no pairs of papers left, and ii) if the final part of the curve trend is too steep (i.e. the tangent at the curve reaches $p^* = 0$ when $U(p^*)$ is negative). In both cases a solution could be to use the smallest p^* for which the tangent at the curve gives a non-negative value of $U(p^*)$.

The value obtained is denoted as $\widetilde{U}_0$, and is used as an approximation of U_0. The procedure used to determine $\widetilde{U}_0$ is sketched in Fig. 4.4(c).

In Fig. 4.4(d)-(e) I report the ranking induced by $\widetilde{U}_0$ respectively for the ten high-level families of PACS codes (panel d) and for the journals published by APS (panel e). It is worth noticing that Electromagnetism and Interdisciplinary Physics are

the two sub-fields with the smallest percentage of missing links, i.e., those in which knowledge flows effectively among articles (and authors), as would be expected if the occurrence of citations were driven by overlaps between topics or research problems. Interestingly, the rate of occurrence of missing citations in Physical Review C ($\widetilde{U}_0 \simeq 0.27$) is almost nine times as large as the one observed in Physical Review Letters ($\widetilde{U}_0 \simeq 0.03$), which is the APS journal with the widest visibility and largest impact[1].

[1]Physical Review Letter has the highest impact factor (8.46 in December 2017) among the other journals of physics published in APS, in which impact factors oscillate between 2.37 and 4.557. A correlation analysis between the journals' ranking proposed and the their impact factors would be interesting to study for a wider selection of journals that have a more homogeneously distributed impact factor, which is not the case in APS journals.

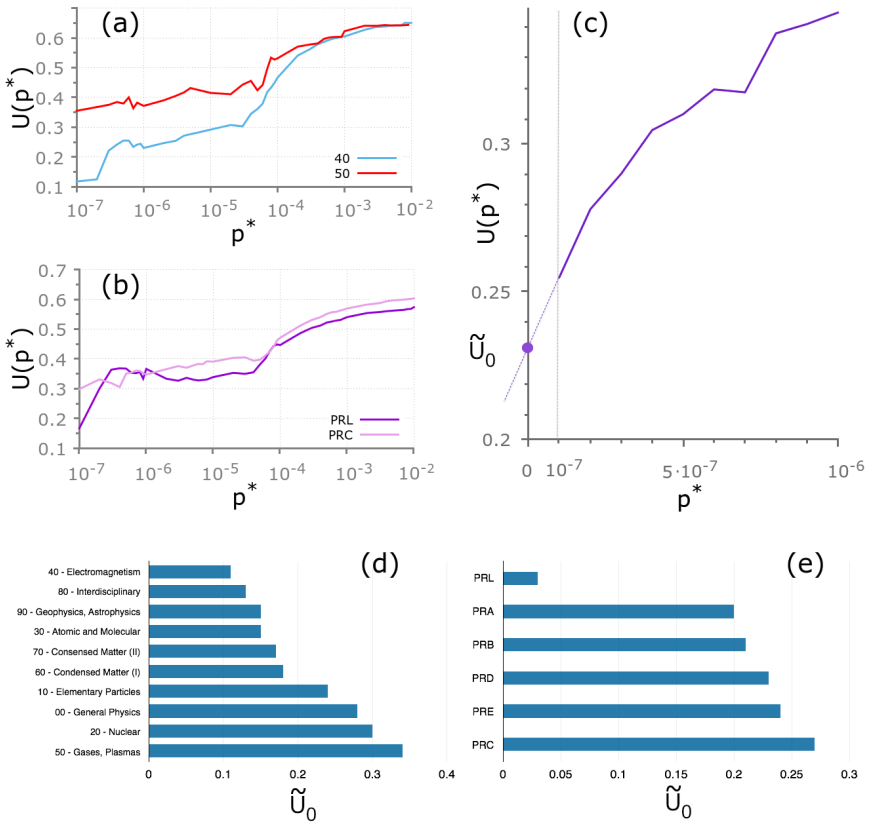

Figure 4.4: **Ranking journals and sub-fields by lack of knowledge flows.** The analysis of missing links restricted to specific sub-fields of physics or single APS journals confirms that the tendency of a citation to occur between a pair of articles increases with the similarity between the bibliographies of the two articles. Panels (a)-(b) show the plots of $U(p^*) = 1 - P_{i \to j}(p^*)$ for different sub-graphs corresponding to (a) two families of PACS codes, namely 40 (electromagnetism) and 50 (Gases and Plasmas), and (b) two APS journals, namely Physical Review Letters and Physical Review C. In panel (c) I sketch the procedure adopted to compute the estimate $\widetilde{U}_0$: I consider the line tangent to the curve $U(p^*)$ at the smallest value of the statistical threshold p^* for which I still have a relatively substantial number of validated pairs (in this case, $p^* = 10^{-7}$), and I define $\widetilde{U}_0$ as the value of the intercept at $p^* = 0$ of that line. In panels (d) and (e) I show, respectively, the rankings of sub-fields and APS journals based on the values of $\widetilde{U}_0$. Notice that Electromagnetism and Interdisciplinary physics are the two sub-fields with the smallest percentage of missing links, i.e., those in which knowledge among articles flows effectively and as would be expected if citations were driven by overlaps between topics or research problems. Interestingly, the lack of knowledge flows between articles published in Physical Review C ($\widetilde{U}_0 \simeq 0.27$) is almost nine times as large as the one identified in Physical Review Letters ($\widetilde{U}_0 \simeq 0.03$), which is the APS journal with the widest visibility and largest impact.

4.5 Conclusions

In this chapter I have proposed a novel method to quantify the similarity between articles based on their bibliographies. The identification of a statistically significant similarity between articles proposed in this chapter can be used to uncover potentially interesting or relevant references that are missing from their bibliographies. This method can thus assist the authors of scientific papers in compiling a list of relevant references, or the editors and reviewers of scientific journals in suggesting otherwise neglected references to the authors of manuscripts submitted for publication. Moreover, public preprint repositories, such as arXiv.org, could automatically quantify the similarity between the bibliography of a newly posted paper and the bibliographies of all other papers in their data set, and then propose a list of papers that the authors might find relevant to their work. The implementation of a recommendation procedure based on statistically significant overlaps between bibliographies might also facilitate the dissemination of scientific results within a scientific field. Problems such as the one shown in Fig. 4.3 can be aptly overcome through the use of this method that enables missing and relevant references to be promptly identified.

Notice that when similarity is evaluated between any two articles published in two different years, all the articles published in the time interval between these two years can only be cited by the more recent article. In principle, it would be possible to modify the method in such a way that the evaluation of similarity would be based only on articles published before the earlier one. However in this chapter I opted not to take the difference in publication years into account in this similarity measure, because this enables pairs of articles published in different years to be more dissimilar than articles published at the same time, all else being equal. This would result from different opportunities, research directions and resources provided by the different time frames in which the two articles were published. This method does indeed capture this time-induced dissimilarity between articles. Moreover, since the analysis was based on the APS data set, the evaluation of the similarity between any two articles was restricted to the overlap between the citations the two articles made only to other articles published in the APS journals. The assessment of similarity could not therefore reflect the entire bibliographies of the two articles.

This limitation can be easily overcome through further analysis of other citation networks extracted from different data sets, such as ISI Web Of Science, or arXiv.org. Finally, this framework can be extended beyond the domain of citations between academic papers, and used for uncovering missing and potentially relevant links in other citation networks, such as those between patents [123, 124] or between the US Supreme Court verdicts [113, 125, 126].

Chapter 5

Interdisciplinarity and homophily in collaboration networks

Recently, scholars across several disciplines have started to analyse and study scientists' scientific performances based on their publications and their ability to win grants or scientific prizes. This domain of research is known today as the "science of science". This chapter will be devoted to study the evolution of physicists' interests over time and in understanding the forces that drives scientific collaborations.

In Chapter 4, I have analysed a network of citations constructed from the American Physical Society (APS) dataset. In this chapter, I will analyse the same dataset but under different perspective. In fact, I will focus on the collaboration between authors. I am interested in understanding the forces that bring two scientists to collaborate. Both homophily (the principle that similarity breeds connection) and heterophily (the principle that connections are created between dissimilar people) have been documented to play an important role in forging connections among individuals. Specifically, in this chapter I will test the interplay between these two mechanisms in the domain of scientific collaboration. Because both principles rely on a measure of similarity, I will propose a measure of "scientific similarity" among scientists, i.e., a similarity based on the expertise and scientific production of each scientist.

This similarity measure will rely on scientists' research interests and expertises. The APS dataset allows me to associate with each author a vector of topics which

reflects the area of physics in which each scholar has published. This is possible because in APS each paper is labelled with at least one, and up to four, codes which reflect the topics treated by the paper. These codes are defined in the Physics and Astronomy Classification Scheme (PACS) and are represented by six digits. By considering the classification's highest hierarchical level (i.e., the decade in which the first two digits of the code fall), I will define ten distinctive areas of modern physics. For each author I will generate a 10-dimensional vector of topics for different time windows of same length, the elements of which will represent the author's contribution to each of the ten defined areas of physics.

The first part of this chapter will be dedicated to characterizing the evolution of physics over time. In fact, the vector of topics will allow me to quantify the temporal evolution of a single author's career by looking at how his or her vector changes over time. In doing so, I will propose a method to quantify the tendency of a physicist to change his or her career towards specialisation or interdisciplinarity. Restricting the analysis to a subset of 54 highly productive authors that have consistently published between 1990 and 2007, I will study the career trajectory of each single author. Results show that the career evolutions towards interdisciplinarity and specialisation are represented in equal proportion. In a second analysis, I will enlarge the set of authors to include those who have published at least five papers in the sliding fixed length time window of five years, from 1990 to 2007. This aggregate analysis over all authors and all area of physics shows that research in physics is evolving towards interdisciplinarity.

In the second part of this chapter, I will study in which measure scientific collaborations are related to scientific similarities between authors. Given a fixed length time window of five years, I will create a collaboration network where two authors are connected if they have co-authored at least one paper in that time window. I will then define a measure to evaluate the scientific similarity between any two authors based on the overlap of the author's topic vectors, and I will investigate whether there is a positive correlation between scientific similarity and the presence of collaboration between authors. Results show a positive monotonic trend which indicates that the more two authors are scientifically similar, the more likely it is that a collaboration between them will be found. This is true up to a given threshold, above which the effect reverses. I will put forward the hypothesis that

scientists with highly similar expertise are less likely to provide each other with the knowledge that they are looking for. As a consequence, scientists will redirect their attention towards colleagues that are less scientifically similar, and therefore more likely to posses a resource (or knowledge) that they may need. In the final part of the chapter, I will propose a network model able to reproduce this non-linear trend.

5.1 Introduction

5.1.1 Authors' career evolution

Scientists' careers follow different trajectories. Among them, there are those who prefer to spend their career on one subject or scientific area; those who progressively shift from one topic to another; those who start with a wide range of interests and end up focusing on one single subject (or vice versa); and those who embrace interdisciplinarity throughout their career. Recently, the study of scientists' careers has become a popular scientific topic known as the "science of science" [127–130]. One of the reasons why this topic has started to attract the interest of many scholars is because the study of research trends plays a major role in the orientation and efficiency of scientific discoveries [131]. Simultaneously, research trends have a strong role in the faculty job market, with major implications on faculties hiring systems or funding allocations [132, 133]. With the study of historical career paths, it is even possible to identify patterns of scientific success based on the evolution of knowledge and interests over time [128, 129].

Different factors may influence the choice of a scientist's career. However, quantitative frameworks that offer an overview of the general research trend of a specific scientific domain remain limited. A first attempt to characterize the evolution of research interests of individual physicists' careers in macro physics areas has only recently been proposed [134].

In this chapter I focus on the scientific domain of physics. I will describe a methodology that provides a description of a scientist's scientific interests in a given time window and I propose a quantification of the extent of change of the physicist's career based on his or her scientific productivity. I will then extend the analysis to the overall trend of all physicists. This will make it possible to understand if the

community of physicists is evolving towards a more interdisciplinary approach or towards a single specialization.

5.1.2 Tie creation mechanisms

Why we befriend, collaborate, or interact with someone is a deep and complicated question to answer. Over the last century, many sociologists have tried to explain the social mechanisms that promote and encourage people to connect with one another. In almost all the mechanisms proposed, social networks have been central to understand and explain the evolution of relationships among social actors.

One of the most elementary but fundamental mechanisms in explaining the creation of a single social tie, has been proposed by James Davis [37] and goes under the name of *triadic closure*. Davis formalizes the idea that there are analogies between clustering and connectedness in a social network. His argument is simple: in a triplet of connected nodes, such that node i is connected to node j and node l, it is more likely that nodes j and l are also connected with each other than would be the case if nodes j and l did not share the connection with i.

In a similar vein, the American sociologist James Samuel Coleman put forward the idea of social embeddedness in which individuals tend to cluster into tightly knit groups that are rich in third-party social relationship [41]. Coleman suggests that the reason behind the creation of these clusters arises from the idea that a specific form of capital, i.e., *social capital*, originates from social structures. The social relationships of these clusters are shaped by the exchange of social capital among individuals. In order to explain the meaning of social capital, Coleman distinguishes three different typologies of capital: human, physical, and social. Human capital is directly related to the skills, knowledge, and know-how of the single individual. Physical capital is represented by the material and personal goods of the individual; and social capital results from a combination of physical and human capital in which the underlying structure of the social network plays a central role. The onset of social relations among individuals is driven by reciprocity, trust, and cooperation and the overall behaviour is primary led towards a common good. In fact, social capital is directly linked to the investment in the social network in which the capital resides. In this way the outcome of a social exchange will be beneficial not only for the

single individual, but it will influence all people taking part in that particular social structure.

Another well documented network growth mechanism is *cumulative advantage*. This principle also goes under the name of the "Matthew effect", or the "rich get richer, the poor get poorer" [42]. This mechanism is based on the hypothesis that once an individual gains a small advantage over other individuals, that advantage will grow over time into a larger advantage. The usage of this term usually refers to issues of popularity or prestige. From a network point of view, it is possible to correlate the fame or social status of an individual with the number of connections (friendships, followers, etc...) that the individual possesses. This effect has been widely investigated in many empirical domains. More recently, in the complex network domain, the famous Barabasi-Albert model [14] has been proposed in order to reproduce the typical power-law degree distribution of nodes in complex networks. Barabasi and Albert argued that the scale-free nature of real networks is rooted in two generic mechanisms: growth and preferential attachment. For what concerns the growth, most real-world networks describe open systems that grow by the continuous addition of new nodes; the preferential attachment is a mechanism which tries to replicate the behaviour of a complex network, such as a social one: the likelihood of a node to connect to another node in the network depends on the node's degree. This principle reproduces the effect of cumulative advantage: nodes that have many connections are more likely to receive more connections over time while nodes with few connections are less likely to grow (in terms of their degree).

5.1.3 Homophily

Among these network mechanisms, a key role is played by homophily, the principle that similarity breeds connection (better known as the common saying "birds of a feather flock together") [18, 43]. Homophily has been empirically documented in a variety of domains, including marriage, friendship, work advice, support, and information transfer [43]. McPherson distinguishes between two different forms of homophily: *status homophily*, based on the similarity of socio-demographic characteristics, and *value homophily*, based on the convergence of similar ideas, beliefs, or mental attitudes among people [43].

A direct connection between the work performed in Chapter 2 and homophily is explained as follows: in addition to clique generation, the literature has long regarded assortative mixing by degree as a possible implication of homophily [44]. In fact, if the creation of a connection is based on the principle that similar individuals are more likely to connect than dissimilar ones, then the resulting network will partition into a number of heterogeneous communities, each composed of similar individuals that share similar ideas, beliefs, or interests. Individuals are expected to forge most of their links with other individuals within their own community, while only a minority of links will be established with other people that belong to different communities. Because the number of social connections of a single individual are constrained by the size of the community to which he or she belongs to, the presence of a community structure implicitly implies that the average number of connections of an individual's nearest neighbours is likely to be correlated with the individual's ones. This leads to assortativity. Thus, because the network community structure is a key factor in determining assortativity, and because one of the main underpinning determinants of the emergence of communities is homophily, if we want to properly understand how and why assortativity occurs in social networks, we need to redirect our focus on homophily.

While homophily boasts a long intellectual tradition in the social sciences, organizational ecologists and economists have simultaneously suggested that similarity can also lead to competition for scarce resources. According to organization theory, similarity and competition go hand-in-hand: high concentration of similar organizations can lead to competition for scarce resources [51–54].

5.1.4 Heterophily

There is a strand of literature that has studied the importance of forging relationships within dissimilar people: the theory of heterophily. One of the leading exponents of the heterophily theory has been the sociologist Georg Simmel [47] with his sociological theory centred on the figure of the "stranger". The stranger is an individual that is not aware of his role in society. He or she is physically present with the people but his or her mind is not with them, but far away. The role of the stranger is to bring innovation, and information among the groups with which

he or she is connected. The reason why people may be more likely to connect with someone who is dissimilar to them could lie in the fact that this connection may help them to gather new information or needed resources (which otherwise might not be found). The strength and importance of the stranger are in his or her *weak ties*: via these ties he or she can canalize and spread information more efficiently. The strength of social ties has been deeply analysed by Mark Granovetter in his highly cited paper [28]. Granovetter defines the strength of a connection through the investment of emotion, time, and intimacy that two individuals put into their (reciprocated) relationship. He categorizes social ties in three ways: strong, weak, and absent. A strong tie is the kind of relationship that connects an individual with his or her siblings, parents, and friends with which he or she spends most of his/her time. An absent tie, as suggested by the word itself, is related to the absence of social connection between any two people. A weak tie can be interpreted as an acquaintances, rather than a real friendship, between any two people.

The presence of weak ties in a social network is essential for information flow between communities: a weak tie often takes the role of bridging social communities, i.e., it corresponds to a connection "*which provides the only path between two points* ([135] *p.198*)" of a network. A weak tie is often a means to spread innovation, as innovators are often seen to belong to the margins of society. To better understand this, imagine the following situation. An actor A needs knowledge that he is not able to reach among his or her friends. If B has a strong connection with A, it is because of the strong overlap of features between the two (homophily), and simultaneously, it will be reasonable to think that B's clique is the same as A's. Thus, B's friends can offer to A only similar resources which end up being unhelpful to A. The strong tie with B gives only minor new resources to draw from. But if A and C share a weak tie, many of C's friends are likely to be strangers (in terms of Simmel's definition) to A, and are more likely to posses needed resources for actor A, inasmuch the overlap in terms of shared knowledge will be as little as possible. This example can help to understand and justify why it is not unlikely to find a connection between people that differ from one another.

5.1.5 The hypothesis

In this Chapter, I want to advance and test the hypothesis that homophily and heterophily are mechanisms that go hand-in-hand in affecting tie creation: if two individuals are similar they are likely to interact (homophily) but they also choose to connect if the social interaction serves their needs (social dependence). That is, if they are sufficiently diverse so as to possess, and be able to exchange, the resources needed to satisfy their objectives (heterophily), a connection is then forged. Homophily would be more likely to impact on a cultural level, based on socio-demographic characteristics (such as gender, ethnicity, age etc.) while heterophily would be more likely to impact at a resource level (being material or intangible ones, such as information or know-how) that each individual possesses and may exchange. However, in my analysis and model, these two mechanisms will act on the same level, with resources and cultural traits representing the same object.

Considering the presence of social dependence among actors, the effects of homophily on tie creation may as a result be mitigated or even reversed. Therefore my aim is to understand and quantify the degree to which social dependence interacts with homophily to affect the way social relationships are created over time. In particular, I shall investigate whether there are non-linear effects of increasing degrees of similarity on the probability of tie creation in collaboration networks. Finally, in the last section of this chapter I propose a model of network creation that combines and extends theoretical arguments of homophily and social dependence.

5.2 The evolution of physics over the years

I start this section with the characterization of the career of a single author. This will be followed by an analysis of the relationships between physicists in Section 5.3. The career characterization is central to define key concepts which I will use for the similarity measure that I will propose in order to evaluate the presence of homophily and heterophily in the collaboration between physicists.

5.2.1 The Dataset

The following study draws on the publicly available database on scientific collaborations of the American Physical Society (APS) journals. The analysis focuses on physicists who published and co-authored on the scientific papers within the journals of the APS. Data extends over 34 years (1980-2014), with 136, 871 authors and 380, 913 articles. For each paper I have access to the authors name and the year of publication. Each article published after 1980 is associated with at least one, and up to four codes defined by the Physics and Astronomy Classification Scheme (PACS) which is a hierarchical partitioning of disciplinary fields within physics based on six characters label. PACS codes consist of ten macro categories which split up into two further levels (for a description of the PACS codes macro categories see the Table in Chapter 4). As an example, the PACS code "81.05.uf" belongs to the broader category of *Interdisciplinary physics* (PACS 80-89), which is part of the sub category *Materials science* (PACS 81) focused on research on *Graphite*. Authors that publish in APS must assign PACS codes based on the list provided on the APS website and these are then approved by reviewers and the editorial office during the revision process. I restricted the analysis only to articles associated with PACS codes, i.e., published after 1980. I have also filtered out all the articles authored by more than 10 co-authors, typically resulting from large experiments in particle physics and high-energy physics.

5.2.2 The vector of topics

The following analysis is based on the reasonable assumption that a paper is the reflection of a scientist's expertise in a given field. Therefore, it is reasonable to expect that each author develops his or her expertise in one or more fields of physics based on the papers that he or she has published. In order to evaluate the physics fields in which an author has published, I look at the PACS codes associated with the papers that he or she has authored. One way to define an author's expertise is to look at the cumulative number of PACS codes in a given time-window. The first two digits of a PACS code are represented by a number that ranges between 00 to 99. I define the set of all PACS codes p such that p corresponds to the collection of

PACS that fall in a given decade[1] with $\mathscr{P} = \{p_1, p_2, ..., p_{10}\}$.

The $\mathscr{P}$ set represents a broad categorization of physics research areas. I quantify the specialization of an author a in a specific domain of physics through the PACS codes p associated with the authored papers within a time window t as

$$P_t^a(p) = \frac{m_t^a(p)}{\sum_{p=0}^{10} m_t^a(p)}$$

where $m_t^a(p)$ represents the author a's multiplicity of PACS codes p, i.e., the number of PACS codes p cumulated in the time window t. The authors' specialization $P_t^a(p)$ represents a way to quantify the expertise of an author in a physics field. For example, the author a has published in a time window t a set of papers associated with the following PACS codes $\{12, 13, 13, 55, 89\}_t$. The multiplicity $m_t^a(p)$ for each set of PACS code p will then be: $m_t^a(p = 10) = 3$, $m_t^a(p = 50) = 1$, $m_t^a(p = 80) = 1$, and $m_t^a(p) = 0$ otherwise. Finally, a's specializations in the time window t are quantified as follow $P_t^a(10) = \frac{3}{5}$, $P_t^a(50) = \frac{1}{5}$, $P_t^a(80) = \frac{1}{5}$, and $P_t^a(p) = 0$ otherwise.

This procedure generates for each physicist a 10-dimensional vector of topics, the elements of which represent the normalized weighted occurrence in each physics area over a given time window. A one year time window will often be too small to analyse a physicists' career with the APS dataset. In fact, even if a physicist has a high productivity, it is possible that he or she does not only only publish in the APS journals and, as a consequence, it may be that in one year none of his or her papers have been published in APS. Therefore, in order to track the career evolution of a physicist in more detail it would be necessary to use other data sources (which is outside of the scope of analysis done in this chapter). To try to overcome this problem I use a five year time window which will help to: i) collect a sufficient number of papers to give a good representation of an author's production, and ii) cover academic cycles (scientific projects, fellowships, PhD, post-docs, etc..). From this point forward, for each measure, I consider a time window of five years.

Cumulating all authors' topic vectors for a time window t, we can rank the most common topic vectors by counting the number of times the same vector is

[1]For instance, the the collection of PACS codes $p = 20$ refers to the broaden category of *Nuclear Physics* and considers all the PACS codes that belong to the set $p = \{20, 21, ..., 29\}$.

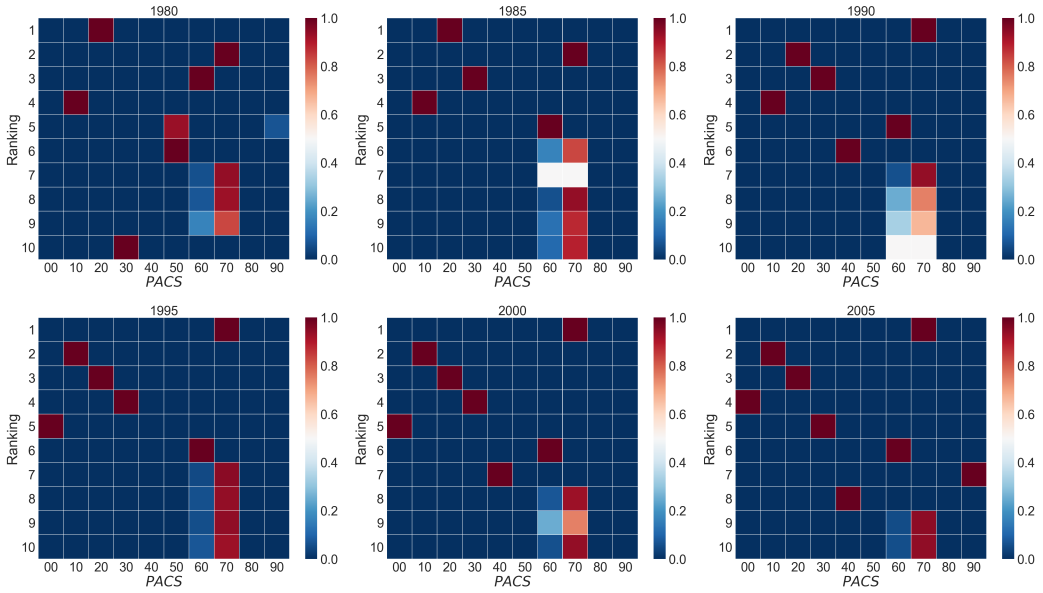

Figure 5.1: **Most frequent topic vectors over time.** Each heatmap represents the 10 most frequent topic vectors among authors for a time window of five years starting with the year written in each title.

repeated across all authors. Fig.5.1 represents six heatmaps that show the 10 most used profiles by physicists that have published in APS across different time windows (1980, 1985, 1990, 1995, 2000, and 2005). The colours go from dark blue ($P_t^a(p) = 0$) to dark red ($P_t^a(p) = 1$). We observe that the 10 most frequent profiles vary over time and the most prevalent attitude is towards the specialization of authors in one area.

5.2.3 Single author's career evolution

In the following study I consider only those authors who have, during the period between 1980-2014, published at least five papers in each time window (i.e., an average of at least one publication per year). Given an author a, I create the career matrix M of a, which is a 10×27 matrix, where 10 represents the number of physics fields and 27 represents the number of time windows from 1980 to 2014. Each column represents the vector of topics. Moving along each row, we move across different time windows. On the left hand side of Fig.5.2 are represented three authors' career matrices through a heatmap representation. This representation helps to highlight

the evolution of an authors' career over time. Moreover, by the use of the Shannon entropy

$$e_t = -\sum_p P_t^a(p) \log(P_t^a(p))$$

it is possible to evaluate the evolution of the career profile of each single author during each time window when plotted over time. A growing entropy trend indicates an author's propensity to spread his or her interests towards a more interdisciplinary career, while a decreasing one reflects an author's propensity to change his or her career towards specialization A flat trend represents an author's fidelity towards the same type of interests over time. On the right hand side of Fig. 5.2 the entropy values calculated for each author over their careers are reported; the straight blue lines represent the linear fit $y = \alpha x + \beta$ of the entropy evolution. The slope coefficient α allows to characterize the careers evolution of each author: a positive coefficient represents the propensity of a scientist to change from a specialized career towards interdisciplinarity, while a negative coefficient reflects the tendency of a scientist to change from an interdisciplinary career towards a more specialised one.

In Fig.5.3 the distribution of all the α slopes calculated for each considered authors is represented. The α coefficient distribution shows the presence of two different trends, with a balance between authors that started their careers as specialized and ended up with an interdisciplinary career and the opposite trend. Only a small percentage of authors remain constant over time.

This method is limited to authors' career evolutions which follow a linear trend. However, the use of overlapping time windows helps smooth the evolution of a scholar's career over time, minimising the effect of non linear evolutions.

Another limitation of the methodology proposed is that it is unable to diversify between trends that are flat over time and that are associated with authors specialised in one field (i.e. specialised authors) in respect to those authors that publish on different areas (i.e. interdisciplinary) and that do not change their specialisations over time. To distinguish between these cases, a different measure that combines both the slope coefficient α (which represents the tendency over time to change) and the intercept β (which may help to understand the level of heterogeneity of interests of an author) could be considered. Both limitations can be considered for future works.

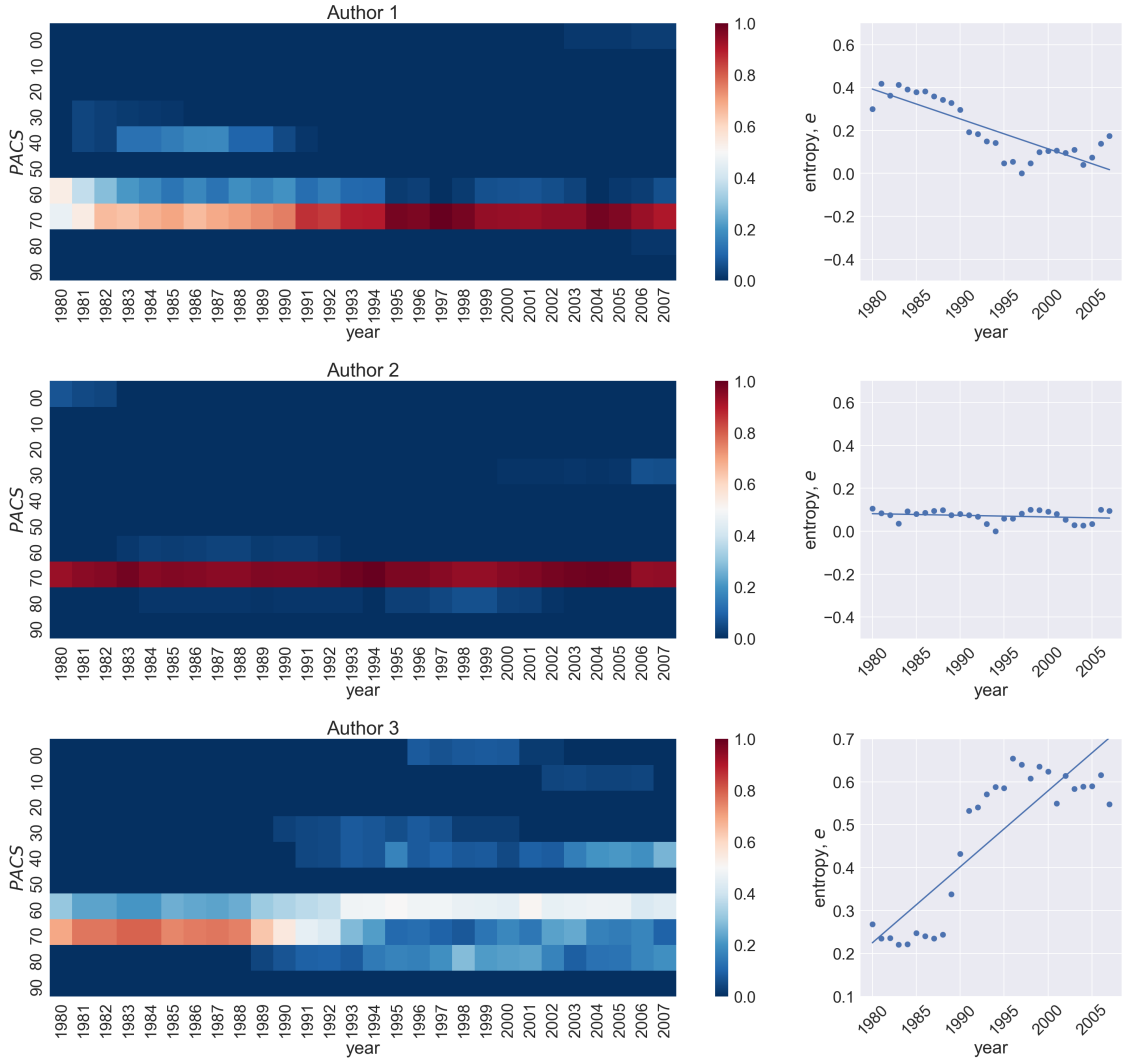

Figure 5.2: **Career evolution.** The first author starts his or her career in an interdisciplinary manner and ends up being more specialized. Author 2 did not change his or her expertise over time. The bottom author starts his or her career in two fields and end up spreading his or her interests. The right hand side graphs represent the entropy trends (blue lines) and their linear fit (green line).

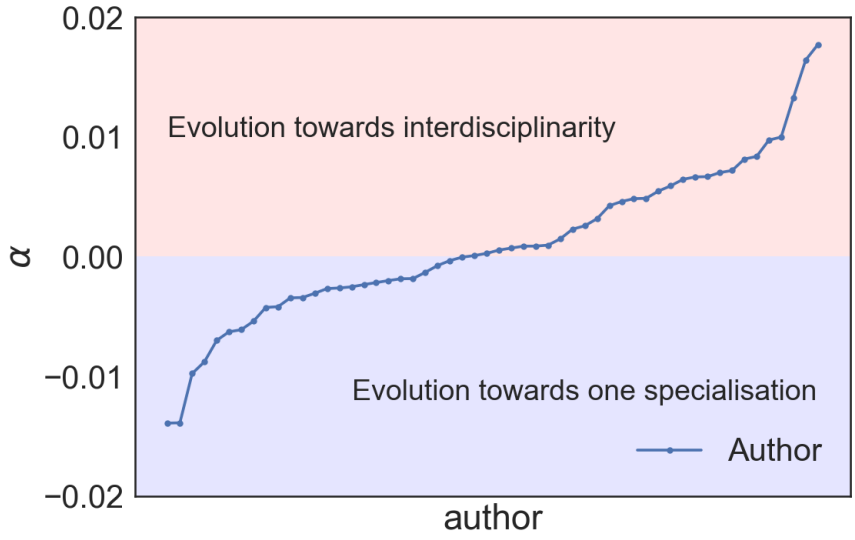

Figure 5.3: **Linear fit coefficient of the entropy trends.** The image shows the different values of the linear fit slope α for the evolution of the entropy trends of each author (see Fig.5.2 green line). We observe two different zones: the red zone, i.e. $\alpha > 0$, populated by authors that start their career as specialized and end up being more interdisciplinary, and the blue zone, i.e. $\alpha < 0$, in which authors start their career with spread interests and end up being more specialized.

5.2.4 Evolution towards interdisciplinarity

In the previous section I studied the career evolution of single authors, and we observed a balance between the two possible career evolutions. The evolution towards either a specialized or an interdisciplinary career appears to be equally preferred by highly productive physicists. By extending the analysis over all authors that have published in APS I can outline the overall trend among physicists.

First, I define the time window t global entropy E as

$$E_t = -\sum_{a=1}^{N_t} \sum_{p} P_t^a(p) \log(P_t^a(p)).$$

This is a measure of all authors' entropies based on their topic vectors in a given time window. A suitable measure to evaluate the entropy evolution over time would then be

$$\langle e_t \rangle = E_t / N_t.$$

The $\langle e_t \rangle$ entropy reflects the averaged entropy of all authors in the time window t. Results are reported in Fig.5.4. The graph shows a positive linear trend, which reflects a general tendency over time for all authors to spread their interests towards more areas, i.e., follow a more interdisciplinary career. See Appendix Chapter 5 for a more in depth analysis.

In the previous section, I was focusing only on highly productive and experienced authors and we observed an almost perfect equilibrium between the two possible career evolutions. However, when we include all the authors, we observe that the overall trend in physics is to follow an interdisciplinary approach. This result shows that physics is an evolving field where authors are increasingly starting to mix expertises in order to produce new and innovative results.

One way to embrace interdisciplinarity for scientists is to study new fields, theories, and methodologies used by other experts outside of their domains. Another way, probably the most used, is by collaborating with experts in other fields. In the next section I will then study the collaboration network among authors that have published in APS.

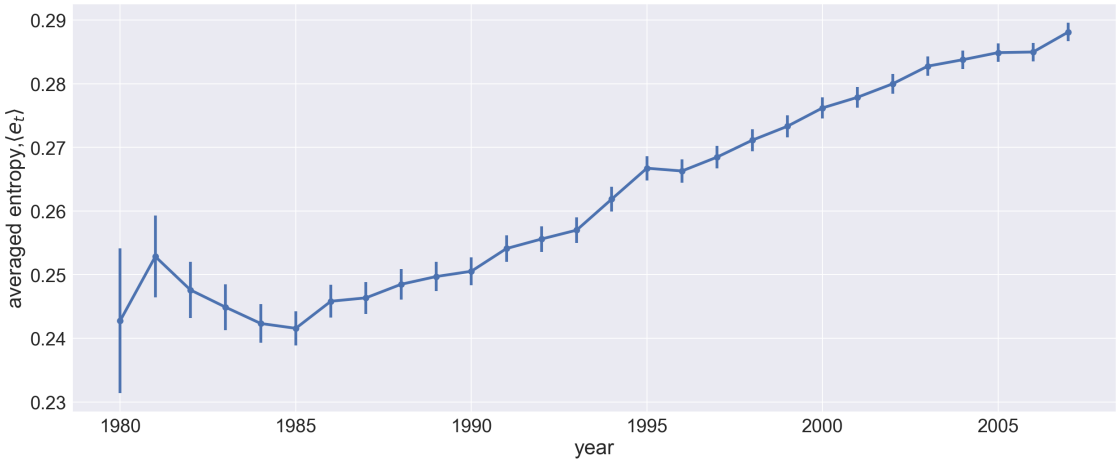

Figure 5.4: **Average entropy.** By evaluating the average entropy $\langle e_t \rangle$ among all authors in a time window t, we observe a positive growing trend that reflects an increasing propensity among physicists towards interdisciplinarity. Standard errors on averages are also reported as bars.

5.3 Evaluating scientific similarity

5.3.1 The collaboration network

I will now focus my attention on the relationships between authors. The suitable way to study the nature of relationships between scholars is through a network approach. In order to create a network of collaboration among physicists from the APS dataset I first collect all articles published over a time window of five years. Second, I produce a two-mode network in which all the authors that have published at least five articles within the considered time window are connected with the articles that they authored. From the two-mode network, I create the one-mode projection in which any two authors are connected with each other if they co-authored at least one paper. The one-mode network represents the collaboration network among physicists that have published in APS journals. This network can be described as a graph $\mathscr{C}(N, E_c)_t$, where N represents the set of authors considered in the time window t, and E_c is the set of collaborations in the time window t. Figure 5.5 gives a graphical representation of how the one-mode projection is obtained from the two-mode network.

Homophily has been largely documented to play a key role in the sociological process of ties creation. If homophily were the only or main social mechanism that explains the presence of collaboration between two scholars, then I would expect to find none or very few pairs of authors that have a high scientific similarity and do not collaborate. Consequently, once the collaboration networks are created for different time windows, I evaluate the scientific similarity of each pair of authors. This similarity measure must be based on the authors' field(s) of expertise, i.e., their vectors of topics.

Knowledge is typically regarded as a resource that accumulates over time [136]. For each pair of authors that have published at least five articles in the time window of five years, I define the scientific similarity $S(a, b)_{t_i}$ of author a in respect to author b as

$$S(a, b)_{t_i} = \sum_{p}^{|\mathscr{P}|} P^a_{t_i}(p)\delta_b(p)$$

where and $\delta_b(p) = 1$ if and only if author b has published at least one paper belonging

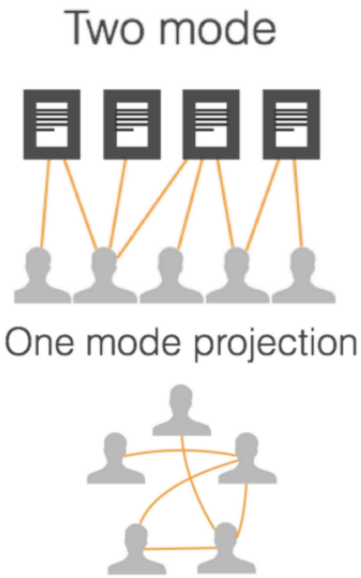

Figure 5.5: **The collaboration network construction.** Scholars that collaborate on the same paper are assumed to be connected in the collaboration network. In doing so, the two- mode network can be transformed into a one-mode projection.

to the PACS category p.

By definition, this measure of scientific similarity is asymmetric, i.e. $S(a,b)_t \neq S(b,a)_t$. To better understand this measure, let's consider the following example in which two authors have published a given number of articles in a time window t_i. I define with A and B respectively the sets of PACS codes of author a and b as $A = \{10, 10, 10, 10, 10, 20, 20, 30\}$ and $B = \{10, 10, 20, 20, 40, 40, 40, 40, 50\}$. Author a is a specialist of domain 10, while author b has focused more in domain 40. Nevertheless, because both authors have worked in domains 10 and 20, but in different proportions, they are similar in a different way. In fact, $S(a,b)_{t_i} = \frac{5+2}{8} = 0.875$ while $S(b,a)_{t_i} = \frac{2+2}{9} = 0.44$. The proposed similarity measure takes into account both the author's "specializations" and the spread of interests of an author in respect to all the physics domains and the possibility to interact with one another and with other authors.

The use of a symmetric measure, such as the Jaccard index, would not be able to quantify the differences in specialization between authors (and would not be able to consider the multiplicity of PACS codes, which is of key importance for the definition

of an author's specialization). For example, if $A = \{10\}$ and $B = \{10, 20, 30, 40\}$, the similarity of the two authors would be in terms of the Jaccard index of $1/4$. It is therefore impossible to highlight any differences between the two authors in terms of knowledge that they could offer to one another or to other authors (that is instead well defined with the proposed measure $S(a, b) = 1$ and $S(b, a) = 1/4$).

5.3.2 Quantifying collaboration in different fields of physics

Given a time window t, I define with $E_{\overline{c}}$ the set of all connections generated by a missing collaboration between any two authors in which their scientific similarity is equal to 1 and that is reciprocated, i.e., $E_{\overline{c}} = \{e := (a, b) | (e \notin E_c) \wedge S(a, b) = S(b, a) = 1\}$ where a and b are two authors.

Given the two sets of edges E_c and $E_{\overline{c}}$, I consider a new network represented by the graph $\mathscr{G}(N, E)_t$ in which $E = \{E_c \cup E_{\overline{c}}\}$. Given this network, it is possible to study the evolution of the collaboration in physics within different fields. I associate a *collaboration coefficient* $C(p)$ with each field in physics similarly to what was done in the previous chapter in order to evaluate the lack of knowledge flow. The score is then based on the ratio between the number of collaborations among authors that have similarity $S = 1$ and the overall number of pairs of authors that have similarity $S = 1$ (which comprises all pairs of authors that have or do not have a collaboration).

Given the graph $\mathscr{G}(N, E)_t$ at a time t, I collect the subset of authors N_p that have published in the physics field p. The generated sub-graph $\mathscr{H}(N_p, E_p)$ describes the network in which $E_p = \{e := (a, b) | e \in E, a, b \in N_p \wedge S(a, b) = S(b, a) = 1\}$. For simplicity, I denote with $E_p^+ = \{e := (a, b) | e \in \{E_p \cap E_c\}\}$ the set of collaborations that took place in the graph $\mathscr{H}$. The collaboration coefficient $C(p)$ for the physics field p at the time window t is then defined as

$$C(p) = \frac{|E_p^+|}{|E_p|}.$$

When $C(p) = 1$, all authors that share the same interests collaborate with one another and, therefore, there are no missing collaborations; when $C(p) = 0$, in the field p there are no collaborations but only "solo" authors that could potentially

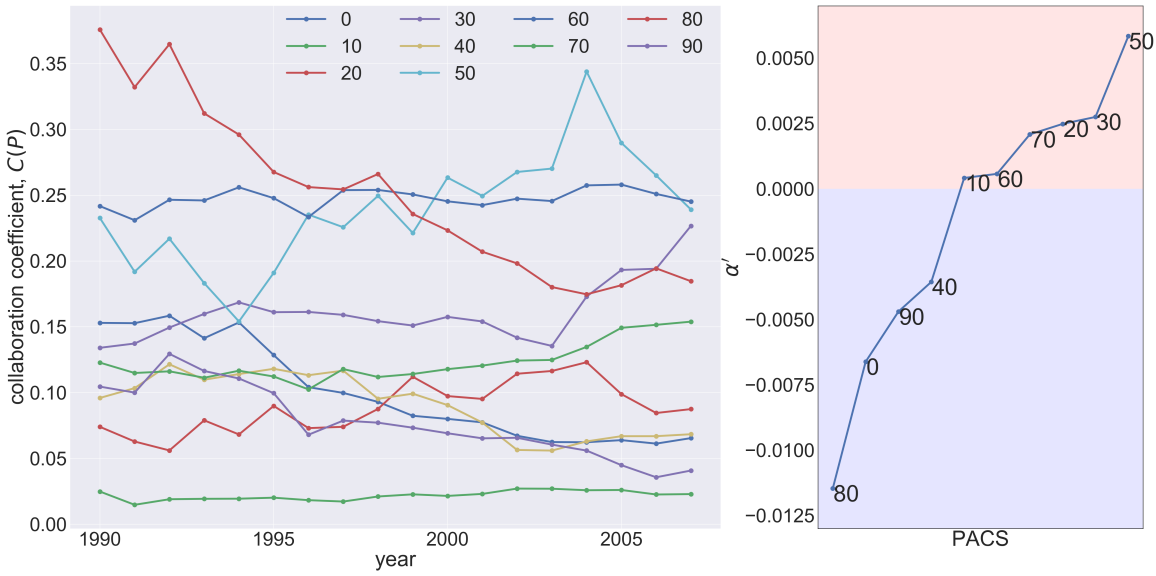

Figure 5.6: **Evolution of the collaboration in physics domains.** The left picture shows the evolution of the collaboration coefficient $C(p)$ for each PACS code p. The picture on the right shows the values of the slope coefficient α' for each linear fit performed on the $C(p)$ curve. The red area corresponds to an increasing collaboration over time, while the blue zone corresponds to a tendency to decrease collaboration within a physics field.

collaborate with similar scholars.

To characterize the overall trend in each physics domain, in Fig.5.6 are reported the evolution in terms of collaborations in all physics fields. Running a linear fit $y = \alpha' x + \beta'$ for each $C(p)_t$ curves, it is possible to highlight two different behaviours: increasing or decreasing collaboration within a physics field. Results are reported in the right hand side picture of Fig.5.6.

The areas that show an increase in terms of collaboration are $p = 20, 30, 70,$ and 50, i.e., respectively *Atomic and Molecular Physics*, *Condensed Matter: Electronic Structure, Electrical, Magnetic, and Optical Properties*. Conversely, the fields $p = 00, 80, 90,$ and 40 which represent *General Physics*, *Interdisciplinary Physics*, *Geophysics, Astronomy, and Astrophysics*, and *Electromagnetism* respectively, show a decrease in their collaboration coefficient. The other fields $P = 10, 60$, i.e., *Physics of Elementary Particles* and *Condensed Matter: Electronic Structure* do not show a significant change over time.

In this analysis I was interested in showing the absence of collaborations between scholars that are completely similar, which goes against the homophily principle and opens new interpretation of the creation of collaboration that I will explain the next sections. A slightly different measure able to create a more sophisticated ranking between fields should take into consideration the number of authors per field. This is out of the scope of the problem treated in this chapter and is left for future works.

5.3.3 Homophily and heterophily: two opposing mechanisms in forging collaboration

Running the similarity measure over time, I showed that many scholars with high similarity do not collaborate. The presence of highly similar scholars without any collaboration may lead us to think that homophily is not the only social mechanism responsible for the creation of social connections in the analysed collaboration networks. However, homophily may encourage the development of a collaboration in a subsequent time window: two authors that have not collaborated in a time window t_i may decide to collaborate in the subsequent time window t_{i+1} because of the overlap of their interests. Fig.5.7 represents an example of this mechanism.

Thus, in order to understand to which degree homophily is a driving force to create future collaborations I shall:

- measure authors' scientific similarity for all time intervals t_i (e.g. 1980 - 1984);
- detect whether each couple has co-authored a paper in the subsequent time window t_{i+1} (e.g. 1985 - 1989).

When I consider the set of authors in the time window t_i and those in the subsequent time window t_{i+1} I measure the similarity between authors that have at least five publications in both time windows. In doing so, nodes in the network at time t_i are the same of the ones of the network at time t_{i+1}.

Choosing only those authors that have at least five publications in both time windows helps to filter out authors that have published only sporadically in APS and enables us to focus on those that have been highly active in both time windows. In doing so, the overall trend arising from the study will be much cleaner and easier

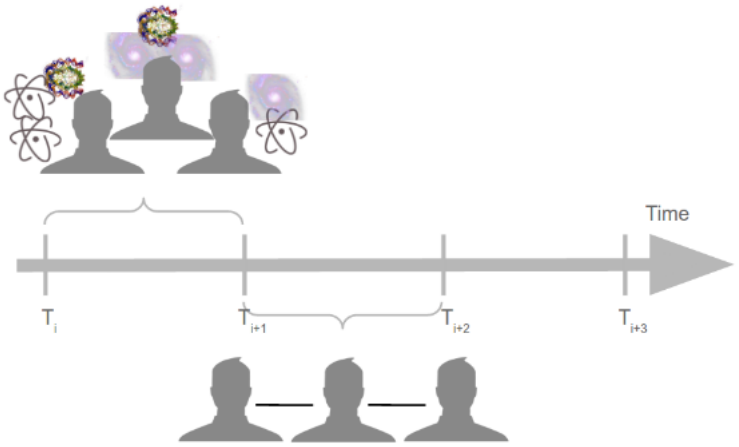

Figure 5.7: **Homophily in fostering future collaborations.** This figure shows how homophily is tested as a mechanism in creating a future collaboration: similarity is measured in the first time window, the presence of a collaboration is evaluated in the subsequent time window.

to interpret. The same study has been performed with a one year and three year time window, obtaining similar results. With a non-time window approach we should take into consideration the inclusion of other effects and the use of other methodologies to explain the creation of a collaboration between two scholars, which is outside of the scope of the analysis performed. Moreover, the choice of using time windows of same length makes it possible to easily compare different time periods over the entire observation period.

In order to avoid strong bias among pairs of scholars that have collaborated at the time window t_i, when I measure the similarity between any two authors that in the previous time window have co-authored a paper, I do not consider the PACS codes arising from papers co-authored by them while measuring their similarity.

Homophily is then tested by evaluating the ratio between the number of collaborations $|E_c(s)|_{t_{i+1}}$ between all pairs of authors associated with a given value of scientific similarity s and the number of all possible pairs $|E(s)|_{t_i}$ with that given value of similarity.

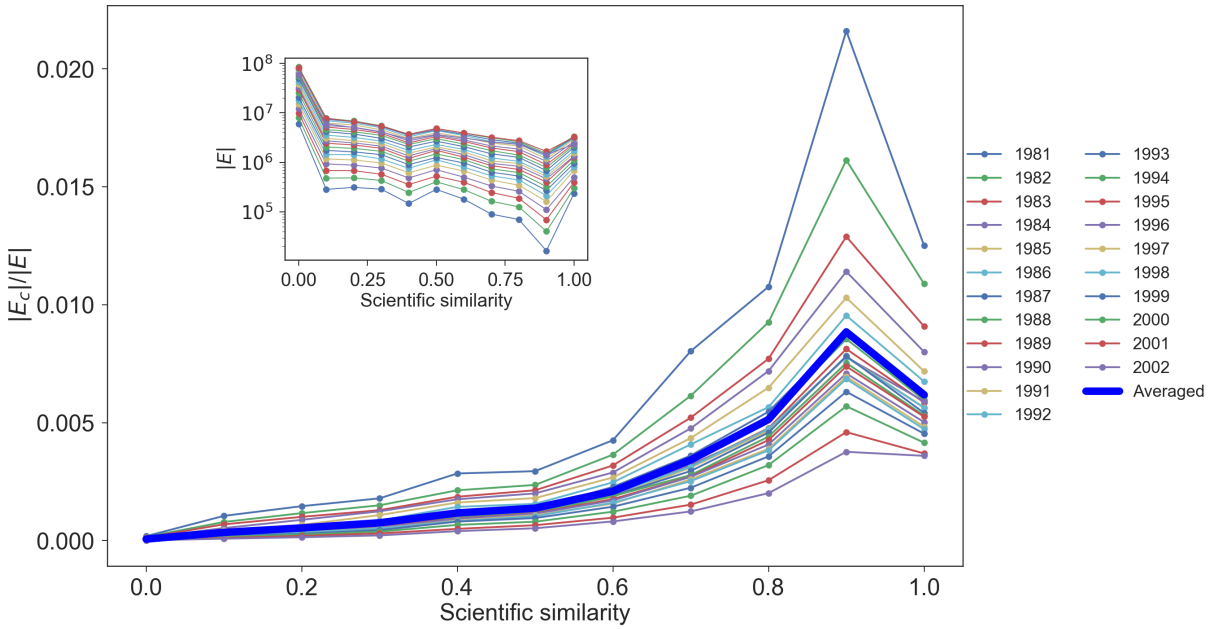

Figure 5.8: **Empirical results.** The figure shows the trend of the ratios between the number of collaborations $|E_c|$ in a time window t_i and the number of pairs of authors $|E|$ with a given similarity at time t_i function of their scientific similarity in the time window t_{i-1}. Curves represent different time windows of 5 years. The reversed "U-shaped" trend is similar in all the represented years. In the inset is pictured the similarity distribution $|E|$ among all authors.

The measure is repeated for all different time windows t_i, in which I calculate the ratio $|E_c|_{t_{i+1}}/|E|_{t_i}$ function of the similarity. By comparing the ratios for different value of similarity s, I obtain a trend of connections based on similarity. If homophily were the only force governing tie creation, we should expect an increasing monotone function reflecting the fact that the probability of tie creation is related to scientific similarity. In other words, authors that are similar are more likely to collaborate with one another than those that are dissimilar.

Empirical results are reported in Fig.5.8. Findings suggest that homophily is a driving force in shaping the relationships among scholars but only up to a certain threshold, beyond which the effects of similarity reverse: if two scholars are too similar, the likelihood of a collaboration decreases. Therefore, it is reasonable to assume that some other concurrent mechanism combines with homophily to affect tie creation. In the next section, I shall propose a model in order to understand

and reproduce the mechanisms that may be responsible for the observed reversed U-shaped effect.

5.4 Modelling the interaction between homophily and heterophily

5.4.1 Axelrod and Centola models

In the vast literature on social interactions and influence, a prominent role has been played by the model introduced by Axelrod [137]. Axelrod's model provides a platform for studying the dynamic interplay between homophily and social dependence, and at the same for understanding how this interplay is also affected by social influence. Just as with the majority of social influence models, each individual is assumed to possess a culture vector C that represents the individual's personal features, where the i-th element of the vector represents the cultural trait of the individual. Each cultural trait in turn takes one of the available q values. Nodes are distributed on a lattice and connected with the neighboring nodes located at the corresponding four cardinal points (N-S-W-E)[2]. The interaction between any two connected nodes happens with a probability proportional to the number of overlapping cultural features divided by the cardinality of C's set. The simultaneous combination of homophily and social influence creates a mechanism that leads to local cultural homogenization. One of the strongest results of Axelrod's work was on the transformation of the overall societal culture. Based on the length of the culture vector C and on the availability of the number of cultural traits q, the synthetic society ends up in two possible scenarios with same length of cultural traits C: on the one hand, when q is small, we face the emergence of a globally polarized social culture, i.e., all the nodes (or the vast majority) share the same components of the culture vector C; on the other hand, when the q is large, the system presents multicultural states with coexistence of different cultural groups.

Another interesting model, based on Axelord's model, has been proposed by Centola *et al.* [138]. Centola's model uses the same previous mechanisms, but it adds a new rule: if two nodes have a zero overlap of their cultural features, then their connection is removed and replaced with the opportunity, made from one of the two considered nodes, to create a new connection with a non-neighboring node.

[2]There is the exception for the nodes that are located at the edges of the lattice that can only have three connections, or two.

With this new rule, both topology and cultural traits evolve over time. Similarly to Axelrod's results, the system evolves either towards a complete homogeneity (in terms of shared cultural traits), or towards multicultural states. However, a main result of the analysis of this model is that, if the system evolves towards multicultural states, the topology of the "society" reflects the cultural separation. In fact, if in Axelrod the network topology cannot vary over time and, we might have two individuals that are completely different in terms of social traits but still share a social connection, in Centola the network evolve into different separated communities in which the individuals share the same culture.

5.4.2 Homophily versus heterophily in Axelrod and Centola's models

By reproducing the two previous models and calculating the ratio between the number of connections and the number of pairs with a given similarity s I will be able to understand if these two models are able to reproduce the previous empirical findings. However, knowing the final states of Axelrod's model, we end up with two possible similarity values: $s = 0$ and $s = 1$. This is due to the fact that the nodes are connected only with nodes that share exactly the same cultural traits ($s = 1$) or with nodes that are completely different ($s = 0$). In Centola's model, if we give enough time to the system to evolve and reach a stationary state, we end up with only one similarity value, i.e., $s = 1$. In fact, nodes are only connected to nodes that share the same cultural traits, but there are no connections with nodes with different cultural traits.

5.4.3 The modified Centola model

In order to reproduce the previous empirical results we can try to use Centola's model and "force" pairs of nodes to break connections when their similarity is equal to 1. In doing so, we are introducing a new process into the connection mechanism. Starting with a random network of $2,500$ nodes, average degree $\langle k \rangle = 4$, and $|C| = 10$, after 10^5 iterations, we obtain the results showed in Fig. 5.9.

At the top of Fig.5.9 are represented the results of the ratio between the number

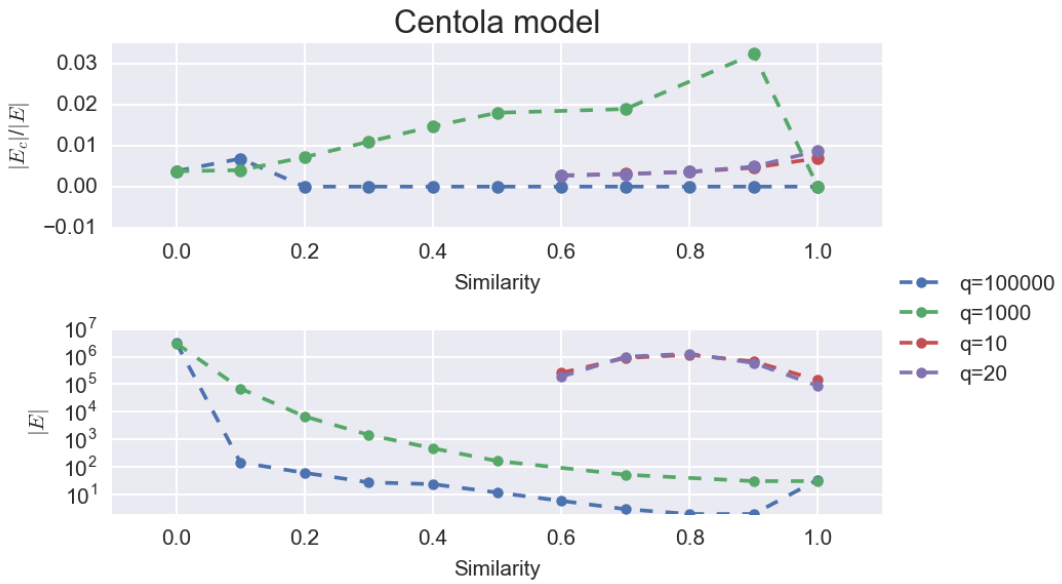

Figure 5.9: **Results on the modified Centola's model**. The top figure shows the trends for the ratio between the number of connected nodes $|E_c|$ and the number of all pairs of nodes function of the similarity for different cultural features q values. The model is the modified Centola model, in which two nodes that have similarity $s = 1$ will break their connection and one of the two nodes looks for a new neighbor. In the bottom figure are plotted the similarity distributions for different values of q.

of connections $|E_c|$ and the number of pairs $|E|$ with a given similarity s. Results show three different trends: for a low number of cultural traits ($q = 10, 20$) there is a growing trend (like the one we would expect in perfect homophily); for $q = 10^5$, there is a reversed U-shaped trend with a maximum at $s = 0.1$; for $q = 10^3$ we have a trend which is similar to the empirical results, i.e., a U-shaped trend with a maximum at $s = 0.9$.

At the bottom of Fig.5.9 are represented the number of nodes with a similarity s. As expected, for small q we have a binomial-like distribution centered in $s = 0.8$. The same distribution appears for larger q but with the maximum shifted to the left at $s = 0$.

Even though for a combination of parameters the model seems to reproduce the empirical findings, a serious drawback of this model is the simulation time. In fact, with a longer simulation time, e.g. after 10^6 iterations, results are similar to what we would expect for the Axelrod model. There are always two values $s = 0$ and

$s = 1$ (except for the case of $q = 10$ in which we have 3 possible values and in the case of $q = 20$, in which we have only one value), as it is shown in Fig. 5.10. Trends are similar for different q values.

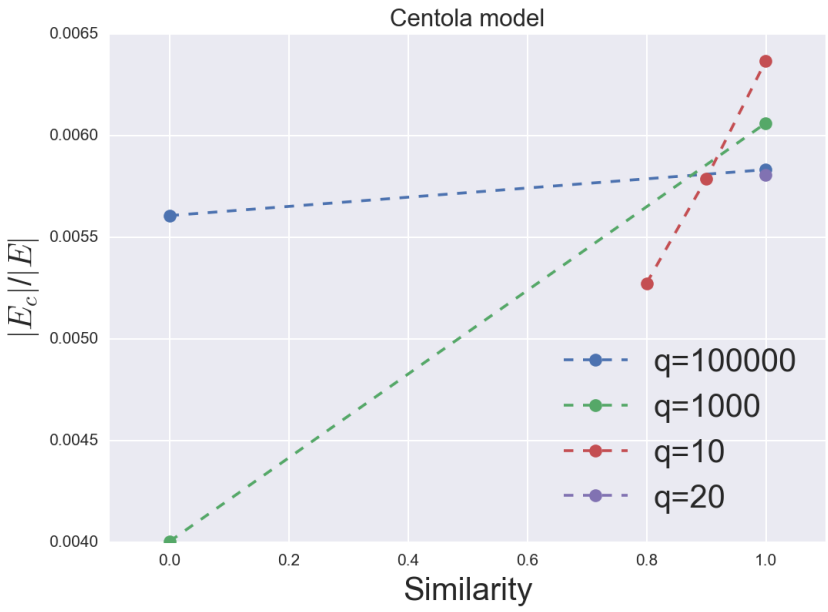

Figure 5.10: **Results on the Centola model with removal for longer simulation time**.

5.4.4 The Homophily and Heterophily (HH) model

I now propose another model which is slightly different from the previous ones. I start with a set of N isolated nodes. This is a major difference between my model and those previously discussed. This model is not initialised with a pre-assigned network.

For simplicity, each node is associated with the coordinates (x, y) that falls on the unit circle[3]. The assigned coordinates represent both the cultural trait and the resource that a node can offer to any other node. I want to model the following mechanism: each node is looking for a different resource (heterophily/social depen-

[3]I could use a similar approach to the Axelrod and Centola models by assigning to each node a cultural vector. I believe that, for the sake of simplicity, the following model description is simpler with the use the unit circle. The transposition to a cultural vector is easily done.

dence) that it can obtain from another node with which they can interact given that they are similar (homophily).

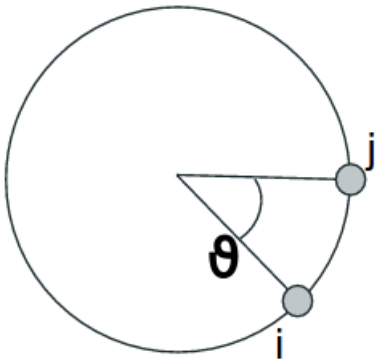

Figure 5.11: **Example of cultural traits association**

In order to reproduce this mechanism, I impose that the probability that a connection is created between two nodes i and j is proportional to their similarity, i.e., the angle between them. The similarity between any two node is a number that fall in the interval $[0, 1]$ evaluated as

$$S(i,j) = \frac{(180 - arcos((x_i \cdot x_j) + (y_i \cdot y_j))\frac{180}{\pi})}{180}$$

where (x_i, y_i) represent the coordinates of node i. Two nodes have $S(i,j) = 1$ if $\theta = 0$ and $S(i,j) = 0$ if $\theta = 180$.

Empirical results show that homophily is the leading mechanism in creating a collaboration. In fact, the reversing trend is observed only for very highly similar scholars. Therefore, the mechanism of tie creation that I propose must strongly rely on homophily and, less frequently, on both homophily and a social dependence mechanism: with a chosen probability p, two nodes connect only if they are similar enough, i.e., $S(i,j) > r_1$ (homophily), otherwise, with a probability $1 - p$ a connection is forged if $S(i,j) > r_1$ and $S(i,j) < r_2$ (the presence of both social dependence/heterophily and homophily are acting simultaneously) where $r_* = rand(0,1)$ and $r_1 < r_2$.

In this model, the property associated with each node represents simultane-

ously a proxy to test the homophilous and the social dependence mechanism. The probability p acts as a "noise" factor, in which the creation of a connection is either triggered only by homophily (or heterophily), or is a combination of both homophily and heterophily. Two conflicting forces are acting concurrently, as in the hypothesis we want to test. The simulations ran over $N = 1,000$ nodes and each node will try to connect with all other nodes (therefore there are $n = 1/2 \cdot N \cdot (N-1)$ attempts of forging a connection).

Finally, once the network is created, I measure for each similarity value s the ratio between the number of connection $|E_c(S = s)|$ and the possible ones $|E(S = s)|$.

If the probability p is absent and a connection is created only when $S(i,j) > r_1$ and $S(i,j) < r_2$ (homophily and heterophily mechanism are balanced), I obtain a reversed U-shaped trend symmetric with a maximum in $s = 0.5$.

Why should we add the probability p in order to obtain results close to the empirical ones? Previous results show that the main force behind a collaboration is homophily. But this is true up to a given threshold, above which an heterophilous effect takes place decreasing the occurrences of tie creation. If the two effects have the same weight, we end up with a "U-shaped" trend in which the curve drops at $S^* = 0.5$. If homophily takes the lead, then the curve will starts to drop at $S^* > 0.5$. The more unbalanced the two effects are in respect to the homophilous (heterophilous) effect, the higher (the lower) the S^* value.

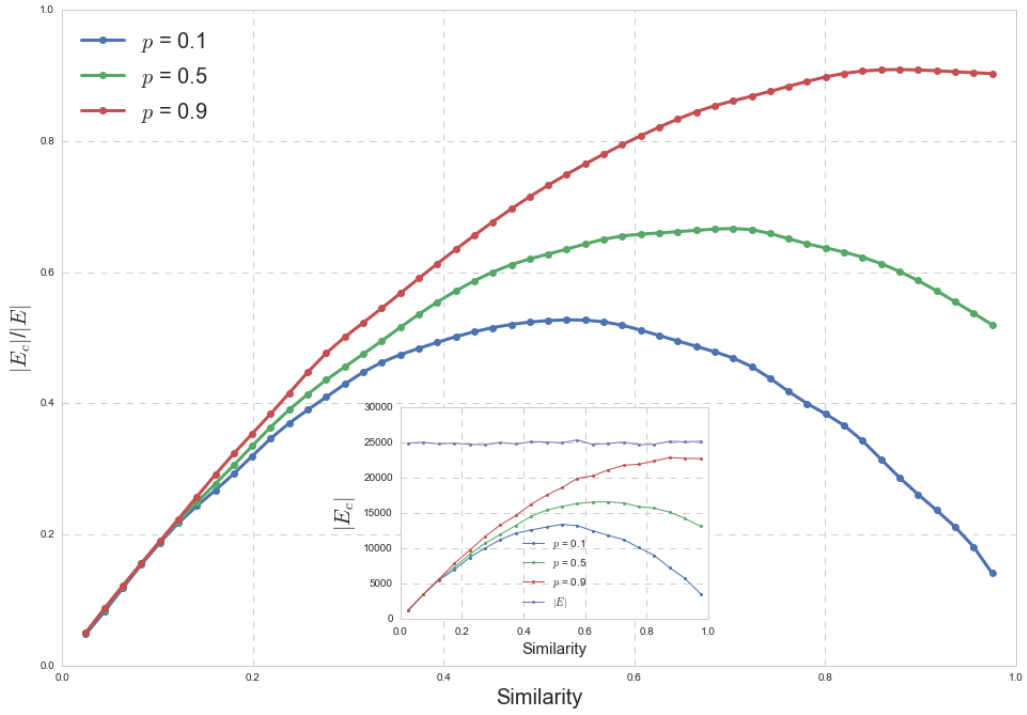

Figure 5.12: **The HH model simulations**. The figure shows the trends for the ratio between the number of connected node $|E_c|$ and the number of all pairs of nodes function of the similarity for different values of the parameter p. In the inset are represented the distributions of $|E_c|$ and the distribution of $|E|$ (grey line).

5.5 Conclusions

In this chapter I have proposed a method to quantify the research interests of physicists and the evolution of their interests over time, and I have analysed the driving forces that forge collaborations among physicists.

For each scientist, I have defined a vector of topics which reflects the area(s) of physics in which they have published over a five year time window. This was possible thanks to the presence of the PACS codes, i.e., alphanumerical labels associated with each paper in the APS dataset analysed. In fact, through the use of PACS codes, I was able to (i) define general domains of physics into well defined sub-domains and (ii) measure the evolution of physicists' interests along their careers. Empirical findings show that, looking at a selection of fifty-four highly productive physicists in the whole observation period of 1980-2013, there are two possible evolutions in a scientist career: one towards specialisation, in which a scientist started his or her career by publishing papers in more than one domain and ended up focusing his or her publications in just one domain, and another towards interdisciplinarity, in which a scientist started his or her career publishing papers in one domain and ended up to spread his or her publications in more than one. This analysis shows a balance between the interdisciplinary and the specialised career evolution when considering the trend of all fifty-four authors. However, by including all authors that have published in a given time window, it becomes clear that there is a general tendency in physics towards an interdisciplinary approach.

The methodology I have proposed to analyse the authors' career evolution shows potential for future work. When I characterise an author's interests into different domains, I associate with him or her a vector which includes the distribution of interests in each area of physics. By doing so, I am assuming that the domain treated by a paper is representative of the author's interests. This is justified by the significant differences between physics areas. However, in a collaborative work, authors may be interested in different aspects and problems of a given domain. Therefore, it would be worthwhile to quantitatively diversify the interests of each author in a single paper on which they collaborate.

The second part of this chapter is devoted to the analysis of the collaboration network among physicists. A number of network growth mechanisms have been sug-

gested to explain how social connections are forged and severed over time. Among them, a key role is played by homophily, the principle that similarity breeds connection. However other studies in the social sciences have pointed in the opposite direction. For example, organizational ecologists have suggested that similarity can lead to competition for scarce resources. According to this research tradition, competition among organizations using similar strategies, of similar size, and in geographical proximity with one another tends to be stronger than competition among dissimilar organizations. I extend the ecological argument to the domain of scientific collaboration and examine the effects that similarity has on tie creation among scientists. My findings suggest that homophily seems to affect tie creation, but only up to a certain threshold, beyond which the effects of similarity reverse. I put forward the hypothesis that actors with high scientific similarity are not likely to provide each other with the resource(s) they are seeking, and therefore they redirect their attention to other less similar collaborators. In order to cast light on these mixed effects of homophily and heterophily on tie creation and to reproduce the empirical findings, I have proposed a model that integrates homophily and social dependence into a unified growth mechanism underpinning the evolution of a social network over time.

To demonstrate the universality of these results, future work should extend the analysis towards other scientific sectors and consider to include scientists' cultural attributes such as gender, ethnicity, age, the geographical location of the institutions, the popularity of the scientific community, and many others. It would be interesting to investigate whether there is any correlation between the particular evolution of the scientific career of a scholar, the way he or she selects his or her collaborators, and the scholar's success measured by standard indicators such as the number of citations received or the author's h-index. These results would be important to understand the key components that have a major impact on the short or long term success of a scientist's career.

Chapter 6

Conclusions and future work

"*I never think of the future - it comes soon enough.*"

— Albert Einstein

"*Science... never solves a problem without creating ten more.*"

— G.B. Shaw

This thesis has dealt with two main research topics. In the first part (Chapters 2 and 3), I have analysed the effects of positive and negative relationships on the topological structure of social networks, focusing on the effect that negative relationships have on the success of organization networks. In the second part (Chapters 4 and 5), I have analysed the presence of homophily as a mechanism behind the creation of citations among scientific papers and its interplay with heterophily in fostering collaboration among scholars.

In Chapter 1, I have carried out a review of the scientific literature in the domains of complexity sciences, social sciences, and organizational sciences. Throughout my thesis I have shown how the theories and methodologies provided by these research domains can be integrated to produce innovative and interdisciplinary approaches to the study of signed networks and tie creation mechanisms.

In Chapter 2, I have conducted a review of how networks are classified based on their topology. In the complex network literature, social networks are considered to

be different from technological or biological ones, not just for the dissimilar intrinsic nature of their nodes, but also for topological reasons. In fact, social networks, i.e., networks composed by individuals and the relationships connecting them, are characterised by the presence of two main properties: (i) a community structure, i.e., groups of nodes in which most of the nodes' connections are shared between one another and (ii) by the presence of a high clustering coefficient, i.e., closed triadic relationships. Both these properties contribute to make the network "assortative", i.e., a network in which nodes are connected to other nodes that share, on average, the same amount of connections. Conversely, biological and technological networks are characterised by the absence of a community structure and by an abundance of open triadic connections among nodes (i.e., a low clustering coefficient). The combination of these two properties contributes to make the network "disassortative", i.e., nodes with many connections are more likely to connect with nodes with few connections and vice versa.

To explain these distinctive properties of social networks, a variety of models from physics, sociology, and computer science domains have been proposed by other scholars: assortative mixing has been related to the underlying community structure of social networks [67], transitivity [69], and homophily [70]. However, the study of mixing patterns by degree has been mainly investigated in unsigned social networks in which nodes are assumed to be connected through positive links [67]. Relatively little attention has been devoted to the emergence of degree correlations in signed networks, and particularly in negative social networks, where individuals are connected through links with a negative connotation, such as distrust, enmity, and competition. In my study, I have focused my attention on the emergence of degree correlations in signed social networks, i.e., social networks in which connections are characterised to possess either a positive or a negative nature (e.g., friendship and enmity, or trust and distrust). Therefore, I have analysed two signed social networks, in which individuals express their trust or distrust toward each other. Empirical findings indicate that negative subnetworks, i.e., the subset of nodes connected only by negative links, are characterized by disassortative patterns, while the overall unsigned network and the positive subnetworks are characterised by an assortative pattern.

In Chapter 3, I have analysed the effects of negative connections in the domain of

organization networks. Specifically, I have studied the effect of competition (negative connections) on both the mobility of employees among start-ups, and on the success of the national ecosystem of start-ups. Drawing on a large dataset of start-ups I have constructed two networks: (i) the network of declared competition, in which a connection from a start-up i towards a start-up j exists if i declares j as its competitor, and (ii) the mobility network, in which a connection from a start-up i towards a start-up j exists if an exchange of employee took place from start-up i to start-up j. Making use of network techniques, I have quantified the effects of competition on the network topology, resulting in a disassortative trend, as seen in the case of social networks with only negative connections. Looking at the overlap between the two start-up networks, the number of resulting connections is very little. This indicates that the exchange of employees between competitors is quite rare. I have then moved my analysis to a national level, aggregating start- ups whose headquarters are based in the same nation. I have defined these sets as national ecosystems. Based on findings related to the effects of competition as an obstructive power for people to move between companies, I have studied the correlation between the success of a national ecosystem and the presence of competition. Empirical findings indicate that these two quantities are anti-correlated, which is interpretable as the more competitive a national ecosystem is, the less successful it is.

In Chapter 4, the study of competition moves from the domain of start-ups to the scientific one. One way to detect competition among scientists is to look for the absence of relevant citation among two scientific papers. In fact, citations in science are an important instrument to affirm the appreciation of a scholar's work. To this end, I have focused on citation networks to cast light on the salience of homophily (namely the principle that similarity breeds connection) for knowledge transfer between papers. Therefore, I have defined the degree to which citations tend to occur between papers that are concerned with seemingly related topics or research problems. Drawing on a large data set of articles published in the American Physical Society (APS) journals, I have proposed a novel method for measuring the similarity between articles through the statistical validation of the overlap between their bibliographies. I have defined the probability $P_{i\to j}(p^*)$ that a citation between any two articles i and j whose similarity is validated at the threshold p^* exists as the ratio between the number of pairs of articles validated at that threshold and the

number of existing citations between those validated pairs. Results suggest that the probability of a citation made from one article to another is indeed an increasing function of the similarity between articles. This study can help to uncover missing citations between pairs of highly related articles, and may thus help identify barriers to effective knowledge flows. By quantifying the proportion of missing citations, I have conducted a comparative assessment of distinct journals and research subfields in terms of their ability to facilitate or impede the dissemination of knowledge. Findings indicate that knowledge transfer seems to be more effectively facilitated by journals of wide visibility, such as Physical Review Letters, than by lower-impact ones.

Chapter 5 has been devoted to the analysis of the same APS dataset with the authors as subject. I have proposed a method to quantify patterns of the evolution of physicists' research interests during their career. First, I have collected for different time windows papers published by each author. Second, for each time window, I have associated a 10-dimensional vector of topics with each author, the elements of which represent the authors productivity in each of the 10 domain of physics defined by codes assigned to papers. Finally, I have analysed the career evolution of each author in terms of the transformation of his or her interests over different domains, looking at the temporal changes of his or her vector of topics. Results indicate that the overall trend in physics is to move towards interdisciplinarity: there is a growing tendency for physicists to spread their interests over more than one domain along their career. In the second part of Chapter 5, I have analysed the creation and evolution over time of collaboration among physicists. Through the definition of a measure of scientific similarity based on the overlap of the authors' vectors of topics, I have proposed a method to associate a collaboration coefficient with each physics area for each time window. I have then studied the temporal evolution of collaboration within a physics field looking at how this coefficient changes over time. Findings show that 60% of physics fields had an increase in terms of collaborations over time and only the 40% had a decrease. Furthermore, I have analysed the correlation between scientific similarity and the collaborations among physicists over time. Findings show a reverse U-shaped trend, which I have interpreted as the combination of two opposing mechanisms: homophily and social dependence. The more two authors are scientifically similar, the more likely it is to find a collaboration

between them. This is true up to a given threshold, above which the effect reverses. Finally, in order to justify this hypothesis I have proposed a network model able to reproduce the non linear trend. The model proposed is an attempt to interpret the way individuals choose to collaborate in a scientific environment based on the combination of the two social mechanisms of homophily and heterophily.

Key contributions

The key contributions brought in this thesis are as follows:

1) In Chapter 2, I have shown that positive and negative relationships differ not only by their intrinsic nature, but also in how they affect the network topology. In fact, the study on two online signed social networks shows a positive correlation between the nodes' degree and their neighbours' degree when sharing a positive relationship, and an anti-correlation in the case of a negative relationship. This is the first time that the presence of anti-correlated degrees in a social network has been connected to the negative relationships shared by individuals. I have also proposed a network model able to reproduce the empirical findings.

2) Through the analysis of the CrunchBase dataset, in Chapter 3, I have analysed the effects of competition on the mobility of employees among start-ups and on the success of the start-up ecosystem of a nation using network approaches. I have created two start-up networks on a worldwide scale: the declared competition network and the mobility network. Empirical findings suggest that (i) competition is an obstructive power for the circulation of people among companies and that (ii) the more competitive a national ecosystem is, the less successful it is.

3) In addition and in accordance with the findings of Chapter 2, the network of declared competition is shown to be disassortative, showing that (i) the type of connection produces effects on a network's topology not only for social networks but also in organization networks, and that, in particular, (ii) negative connections seem to generally produce a disassortative trend.

4) Another contribution from Chapter 3 is related to the use of network techniques in order to extract a hierarchical structure from unstructured data. I have combined two networks techniques (the network backbone analysis and the community detection algorithm) in order to extract meaningful information from the CrunchBase dataset concerning start-ups' market sectors. Characterising each nation with start-up activity into the identified macro industry sectors, allowed me to reveal differences and similarities between nations and distinguish the various patterns of activity of different start-up ecosystems.

5) In Chapter 4, I have analysed the citation network among papers published in the American Physical Society (APS) journals and I have proposed a novel method to identify the absence of statistically relevant citations. Results show that the more two papers share a significant overlap in their bibliographies, the more likely there is to be a citation between them. In other words, homophily has been found to be an important mechanism in citation networks shaping the structure and evolution of knowledge transfer between academic papers. Finally, I have proposed a way of ranking physics areas and journals based on the number of missing citations among their papers.

6) In Chapter 5, I have proposed a method to quantify the research interests of physicists and the evolution of their interests over time. I have then measured the tendency of physicists to change their career towards either interdisciplinarity or a specialisation. Results show that in physics the tendency over the last 30 years is to move towards an interdisciplinary approach.

7) In the second half of Chapter 5, I have put forward and tested the hypothesis that two opposing mechanisms such as homophily and heterophily contribute in forging connections among scholars in the domain of scientific collaboration. Findings suggest that homophily seems to affect tie creation but only up to a given threshold, beyond which the effects of similarity reverse. I have then put forward the hypothesis that highly similar individuals are less likely to provide each other with a resource (material or immaterial, such as knowledge or skills) that the scholar is looking for. As a consequence he or she will redirect his or her attention to other, less scientifically similar collaborators. Finally, I have

proposed a network model based on both homophily and heterophily able to reproduce the empirical findings.

Future works

My work open to future works and has various implications for research.

First, findings on signed social networks can be regarded as a platform for further studies of mixing patterns in complex networks. In fact, my study suggests that the sign of the links could, in principle, be inferred simply from the analysis of the structural properties of a network. From this perspective, findings can help inspire the development of a quantitative measure to uncover the hidden sign of the links from the type of mixing patterns exhibited by a network.

The approaches and methodologies developed in Chapter 3 may become a framework to help build innovative ecosystems, showing that a competitive environment can damage rather than foster the success of the whole national ecosystem. This could be done by helping the free flow of people within national boundaries and by subsidising companies that encourage collaboration and exchange of personnel between companies.

The analysis proposed on citation networks (Chapter 4), on authors' careers evolution, and on scientists' collaboration networks (Chapter 5) could be extended to other scientific domains, using larger dataset such as the ISI Web Of Science. As we witness a continuously increasing production of preprints and publication of new articles, it has become particularly difficult for authors to keep abreast of scientific developments and relevant works related to the domain of interest. As a result, lack of knowledge of prior or current related work and missing relevant citations may occur quite often. The method presented in Chapter 4 can help the scientific community precisely to address this problem. In particular, it can be used not only by authors to integrate the bibliographies of their work, but also by editors of scientific journals to uncover relevant missing citations and identify the appropriate reviewers for the papers they are considering for publication.

Concerning the study of scientists' career evolutions (Chapter 5), it would be of great interest to quantitatively measure the similarities and differences of distinct scientific areas and understand what are the overall trends in each specific domain.

This kind of information could be used by funding bodies to quantitatively evaluate the expertise of the projects' coordinators based on their career paths. Finally, the study on collaboration among physicists should extend to investigating the presence of any correlation between a scholar's success and the way he or she selects his or her collaborators. These results will represent an important step forward to unveil the key components that have a major impact on the short or long term success of a scientist's career.

Appendix Chapter 4

Effect of time on citations

In order to understand the effect of time on the methodology, I report the effect of time (age of the paper) on papers' citations. Specifically, I plot the average age of the citing and the cited papers function of the k in-degree of each S^k subsets.

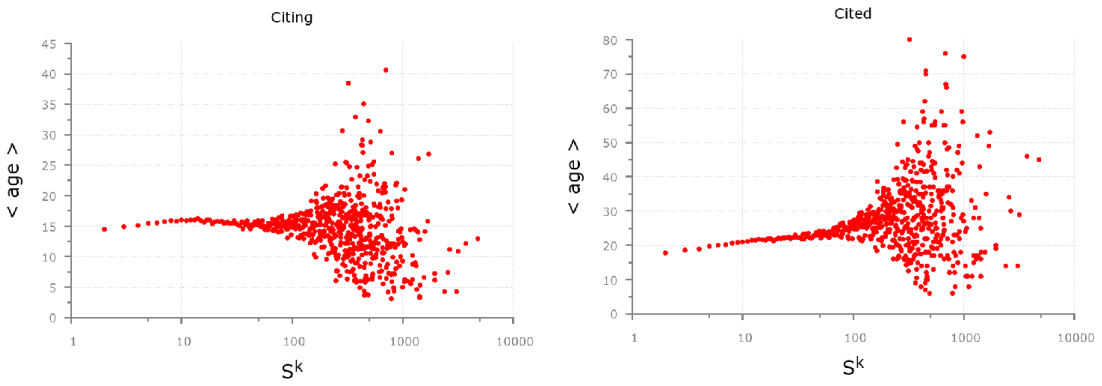

Figure 1: **Papers ages function of the in- and out-degree citations.**

In the left hand side figure we observe that there is no temporal dependence for the citing papers over the different subsets S^k. From the right hand side figure, we observe that the average age is only slightly growing as the number of incoming citations k grows. These two figures show that the subsets S^k do not show any relevant bias in respect to time.

False Discovery Rate (FDR) statistical test

The validation of a given pair (i, j) in the FDR method is performed as follows [122]. We set a statistical threshold p^* and we assume that there are in total N_t tests. Then, the p-values of different tests are first arranged in increasing order ($q_1 < q_2 < ... < q_{N_t}$), and the rescaled threshold is obtained by finding the largest t_{max} such that

$$q_{t_{max}} < \frac{p^* t_{max}}{N_t}, \tag{1}$$

where N_t is the number of tests. In this specific case, N_t is the number of distinct pairs of articles that are tested over all the sets S^k of in-degree classes in the citation network. Then we compare each p-value $q_{ij}(k)$ with the rescaled threshold, and we validate the pair (i, j) if $q_{ij}(k) < p^* \, t_{max}/N_t$.

Appendix Chapter 5

Averaging entropies

The analysis performed in Section 5.2.4 shows a positive linear trend which reflects an average tendency over time for all authors to spread their interests over more than one area of research, thereby following a more interdisciplinary career.

This study is made on the assumption that the distribution average is representative of the entropies distribution. This could be the case if, for example, the entropies are normally distributed. For each time window considered, the entropies follow similar probability density distributions as the one shown in Fig. 2(a). The overall distribution follows a normal distribution except for the pick in $e = 0$, which represents the specific case of specialised authors. Therefore, the resulting entropy distribution seems to show two different behaviours: one for specialised authors, i.e. $e = 0$, and one for "interdisciplinary" authors, i.e. $e \neq 0$.

To better understand the general tendency in physics for authors' pursuit of a specialised or an interdisciplinary career, I can study separately the evolution of the relative percentage of specialised authors over time and the evolution of the average distribution of entropy given for the interdisciplinary authors (without considering the contribution arising from specialised authors). Figure 2(b) shows that the relative percentage of specialised authors is decreasing over time. Results show that the relative number of specialised authors has diminished passing from a 21% in 1980 of the relative population to the 12% in 2007. Based on these results we can argue that, in physics, there is a tendency for authors to become less specialised over time.

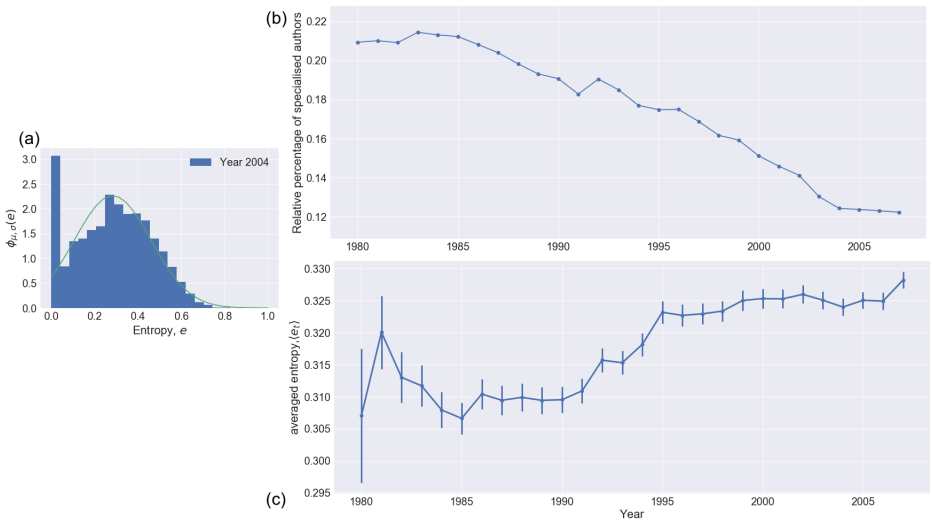

Figure 2: **Topic entropy evolution over time.** **(a)** Entropy distribution of all the authors for the time window of five years starting on 2004. The pick in $e = 0$ represents the authors that have published in only one physics field. The rest of the distribution ($e \in (0, 1]$) follows a normal distribution. **(b)** Relative percentage of specialised authors over time. This figure represents the relative percentage of authors that are specialised ($e = 0$) in each year. We observe a decreasing trend over time that shows a tendency for authors to become less specialised over time. **(c)** Average entropy over time without considering specialised authors.

Moreover, in accordance with what is shown in Section 5.2.4, the average of the entropy distribution over time follows again an increasing trend (Fig.2(c)). This figure differs from Fig.5.4 because I am only considering the average arising from the distribution of the entropies without considering specialised authors.

Presented work

During my PhD, the projects I have worked on have also been presented to various conferences. In particular, the work on *degree correlations in signed social networks* (Chapter 1) has been presented to:

- international conference on network science NetSci14, in Berkeley, California (2014);
- XXXV international Sunbelt social network conference, in Brighton, United Kingdom (2015); and
- Complex Systems Digital Campus 2015, at the World e-Conference.

Second, the work in Chapter 2 on *the nature of competition in start-up ecosystems* has been presented at:

- the cross-disciplinary workshop Data Natives, in London, United Kingdom (2017); and
- the 2017 Pint of Science event in London.

Both presentations have integrated the results and the analysis made also in the submitted paper *Predicting success in the worldwide start-up network*.
Third, the work on *Homophily and missing links in citation networks* (Chapter 3) has been published and presented in:

- the conference on complex systems (CSS16), in Amsterdam, Netherlands (2016);
- the 2nd Imperial College SIAM chapter annual conference, in London, United Kingdom (2016) for which I won the best talk award;

- the cross-disciplinary workshop Data Natives, in London, United Kingdom (2016); and
- has been accepted but, unfortunately, has not been presented at the international Conference on Network Science (NetSci16), at Seoul, Korea (2016).

Fourth, the project on *homophily and tie creation in social networks* presented in Chapter 4 has been presented on different forms in many conferences. In particular:

- the Thirteenth mathematics of networks meeting, at the Imperial College, London, United Kingdom (2014);
- cross-disciplinary workshop, at the London Institute for Mathematical Sciences (LIMS)(2015);
- cross-disciplinary workshop Data Natives, in London, United Kingdom (2015);
- ARS15, Capri, Italy (2015);
- the international conference on network science (NetSci15), in Zaragoza, Spain (2015);
- the international conference on computational social science (ICCSS), in Helsinki, Finland (2015); and
- has been accepted but has not been presented to CompleNet 2016 conference in Dijon, France (2016).

In addition to the projects mentioned in this thesis, I have also work on a project that examines the effects of collaboration among theatres on creativity, economic, and popularity performances. I presented the results of this project to the 2nd annual international conference on computational social science (IC2S2 2016), in Evanston, IL, USA.

References

[1] M. S. Granovetter, "Economic action and social structure: The problem of social embeddedness," *American Journal of Sociology*, vol. 91, no. 3, pp. 481–510, 1985. 1, 2, 60

[2] T. Hobbes, *[1651] Leviathan.* New York, NY: Penguin, 1968. 2

[3] D. Ricardo, *[1816] On the Principles of Political Economy and Taxation.* Cambridge, MA, USA: Cambridge University Press, 1951. 2

[4] D. Wrong, "The oversocialized conception of man in modern sociology," *American Sociological Review*, vol. 26, no. 2, pp. 183–196, 1961. 2

[5] S. Bowles and H. Gintis, *Schooling in Capitalist America.* New York, NY, USA: Basic Books, 1976. 2

[6] T. Parsons, *The Structure of Social Action.* New York, NY, USA: McGraw-Hill, 1937. 2

[7] M. J. Piore, *Notes for a Theory of Labor Market Stratification.* Cambridge, MA, USA: Heat, 1975. 2

[8] F. J. Roethlisberger and W. J. Dickson, *Management and the Worker.* Cambridge, MA, USA: Harvard University Press, 1939. 3

[9] J. Potterat, L. Phillips-Plummer, S. Muth, R. Rothenberg, D. Woodhouse, T. Maldonado-Long, H. Zimmerman, and J. Muth, "Risk network structure in the early epidemic phase of hiv transmission in colorado springs," *Sexually Transmitted Infections*, vol. 78, no. suppl 1, pp. i159–i163, 2002. 3

[10] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002. 3

[11] N. D. Martinez, "Artifacts or attributes? Effects of resolution on the little rock lake food web," *Ecological Monographs*, vol. 61, pp. 367–392, 1991. 3

[12] R. Cohen, K. Erez, D. Ben-Avraham, and S. Havlin, "Resilience of the internet to random breakdowns," *Physical Review Letters*, vol. 85, no. 21, p. 4626, 2000. 3

[13] M. E. Newman, "The structure and function of complex networks," *SIAM review*, vol. 45, no. 2, pp. 167–256, 2003. 4

[14] A. L. Barabási, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, Oct. 1999. 4, 10, 23, 41, 45, 91, 116

[15] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, "Complex networks: Structure and dynamics," *Physics reports*, vol. 424, no. 4, pp. 175–308, 2006. 4

[16] S. Wasserman, *Social Network Analysis: Methods and Applications*, vol. 8. New York, NY, USA: Cambridge University Press, 1994. 4

[17] D. J. Brass and M. E. Burkhardt, *'Centrality and power in organizations.'*, pp. 191–215. In Nohria, Nitin and Eccles, Robert G., eds., *Networks and Organizations: Structure, Form, and Action*, Boston, MA, USA: Harvard Business School Press, 1992. 5

[18] J. Freeman, "The population ecology of organizations," *American Journal of Sociology*, vol. 82, no. 5, pp. 929–964, 1977. 5, 10, 116

[19] W. W. Powell, K. W. Koput, and L. Smith-Doerr, "Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology," *Administrative Science Quarterly*, vol. 41, no. 1, pp. 116–145, 1996. 5, 6

[20] N. Nohria and R. G. Eccles, *Networks and Organizations: Structure, Form and Action.* Boston, MA, USA: Harvard Business School Press, 1992. 5

[21] M. H. Best, *The New Competition: Institutions of Industrial Restructuring*. Cambridge, MA, USA: Harvard University Press, 1990. 5

[22] H. C. White, *Chains of Opportunity*. Cambridge, MA, USA: Harvard University Press, 1970. 5

[23] R. M. Kanter, *The Change Masters*. New York, NY, USA: Simon and Schuster, 1983. 6

[24] S. R. Barley, J. Freeman, and R. C. Hybels, *'Strategic alliances in commercial biotechnology.'*, pp. 311–347. In Nohria, Nitin and Eccles, Robert G., eds., *Networks and Organizations: Structure, Form, and Action*, Boston, MA, USA: Harvard Business School Press, 1992. 6

[25] S. P. Borgatti and P. C. Foster, "The network paradigm in organizational research: A review and typology," *Journal of Management*, vol. 29, no. 6, pp. 991–1013, 2003. 6

[26] N. Lin, *Social Capital*. New York, NY, USA: Cambridge University Press, 2002. 6

[27] R. S. Burt, *Structural Holes: The Social Structure of Competition*. Cambridge, MA, USA: Harvard University Press, 1992. 6, 13

[28] M. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973. 6, 13, 14, 118

[29] W. E. Baker, R. R. Faulkner, and G. A. Fisher, "Hazards of the market: The continuity and dissolution of inter-organizational market relationships," *American Sociological Review*, vol. 63, pp. 147–177, 1998. 6

[30] J. L. Bradach and R. G. Eccles, "Price, authority, and trust: From ideal types to plural forms," *Annual Review of Sociology*, vol. 15, pp. 97–118, 1989. 6

[31] B. Demil and X. Lecocq, "Neither market nor hierarchy nor network: The emergence of bazaar governance," *Organization Studies*, vol. 27, no. 10, pp. 1447–1466, 2006. 6

[32] R. S. Achrol, "Changes in the theory of interorganizational relations in marketing: Toward a network paradigm," *Journal of the Academy of Marketing Science*, vol. 25, no. 1, pp. 56–71, 1997. 6

[33] F. Heider, *The Psychology of Interpersonal Relations*. New York, NY, USA: Wiley, 1958. 7

[34] L. Festinger, *A Theory of Cognitive Dissonance*, vol. 2. Stanford, CA, USA: Stanford University Press, 1957. 7

[35] F. Heider, "Attitudes and cognitive organization," *The Journal of Psychology*, vol. 21, no. 1, pp. 107–112, 1946. 8, 21

[36] D. Cartwright and F. Harary, "Structural balance: a generalization of Heider's theory," *Psychological Review*, vol. 63, no. 5, pp. 277–293, 1956. 8, 21, 24, 35, 43, 49

[37] J. A. Davis, "Clustering and structural balance in graphs," *Human Relations*, vol. 20, pp. 181–187, 1967. 8, 10, 21, 115

[38] D. Brown, "The structuring of Polopa feasting and warfare," *Man*, vol. 14, no. 4, pp. 712–733, 1979. 9

[39] M. W. Young *et al.*, *Fighting With Food. Leadership, Values and Social Control in a Massim Society*. Cambridge, England: Cambridge University Press, 1971. 9

[40] E. O. Laumann, *The Organizational State: Social Choice in National Policy Domains*. Madison, WI, USA: University of Wisconsin Press, 1987. 9

[41] J. S. Coleman, "Social capital in the creation of human capital," *American Journal of Sociology*, vol. 94, pp. 95–121, 1988. 10, 115

[42] S. J. Gould, *The Structure of Evolutionary Theory*. Cambridge, MA, USA: Harvard University Press, 2002. 10, 116

[43] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: homophily in social networks," *Annu Rev Sociol*, vol. 27, 2001. 10, 11, 12, 92, 102, 116

[44] M. E. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, no. 20, p. 208701, 2002. 11, 117

[45] D. Krackhardt and J. Hanson, "Informal networks: The company behind the chart," *Harvard Business Review*, vol. 71, no. 4, pp. 104–111, 1993. 12

[46] M. T. Hansen, "The search-transfer problem: The role of weak ties in sharing knowledge across organization subunits," *Administrative Science Quarterly*, vol. 44, no. 1, pp. 82–111, 1999. 12, 13

[47] G. Simmel, *The Sociology of Georg Simmel*. Glencoe, IL, USA: Free Press, 1950. 12, 14, 60, 117

[48] R. S. Burt, "Structural holes and good ideas," *American Journal of Sociology*, vol. 110, no. 2, pp. 349–399, 2004. 13, 60

[49] R. S. Burt, *Brokerage and Closure: An Introduction to Social Capital*. Oxford: Oxford University Press, 2005. 13, 60

[50] M. Tortoriello, R. Reagans, and B. McEvily, "Bridging the knowledge gap: The influence of strong ties, network cohesion, and network range on the transfer of knowledge between organizational units," *Organization Science*, vol. 23, no. 4, pp. 1024–1039, 2012. 13

[51] G. Ahuja, F. Polidoro, and W. Mitchell, "Structural homophily or social asymmetry? the formation of alliances by poorly embedded firms," *Strategic Management Journal*, vol. 30, no. 9, pp. 941–958, 2009. 15, 117

[52] J. A. Baum and J. V. Singh, "Organizational niches and the dynamics of organizational mortality," *American Journal of Sociology*, vol. 100, pp. 346–380, 1994. 15, 117

[53] M. T. Hannan and J. Freeman, *Organizational Ecology*. Cambridge, MA, USA: Harvard University Press, 1993. 15, 117

[54] J. D. Thompson, *Organizations in Action: Social Science Bases of Administrative Theory*, vol. 1. New York, NY, USA: McGraw-Hill, 2011. 15, 117

[55] A. Grandori, "An organizational assessment of interfirm coordination modes," *Organization Studies*, vol. 18, no. 6, pp. 897–925, 1997. 15

[56] M. Olson, *The Logic of Collective Action*. Cambridge, MA, USA: Harvard University Press, 1965. 15

[57] O. E. Williamson, "The economics of organization: The transaction cost approach," *American Journal of Sociology*, vol. 85, pp. 548–577, 1981. 15, 16

[58] R. M. Emerson, "Power-dependence relations," *American Sociological Review*, vol. 27, pp. 31–41, 1962. 16

[59] P. V. Marsden, "Restricted access in networks and models of power," *American Journal of Sociology*, vol. 88, pp. 686–717, 1983. 16

[60] M. Gargiulo, G. Ertug, and C. Galunic, "The two faces of control: Network closure and individual performance among knowledge workers," *Administrative Science Quarterly*, vol. 54, no. 2, pp. 299–333, 2009. 16

[61] J. Pfeffer, *Power in Organizations*. Marshfield, MA, USA: Pitman, 1981. 16

[62] M. E. Newman and J. Park, "Why social networks are different from other types of networks," *Physical Review E*, vol. 68, no. 3, p. 036122, 2003. 22

[63] R. Albert and A.-L. Barabási, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, p. 47, 2002. 22, 30

[64] P. ErdHos and A. Rényi, "On random graphs," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959. 23

[65] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-world networks," *Nature*, vol. 393, pp. 440–442, 1998. 23

[66] M. Newman, "Assortative mixing in networks," *Physical Review Letters*, vol. 89, p. 208701, Oct. 2002. 23, 29, 30, 35, 41

[67] M. Newman and J. Park, "Why social networks are different from other types of networks," *Physical Review E*, vol. 68, p. 036122, Sept. 2003. 23, 24, 30, 34, 35, 54, 146

[68] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The anatomy of the facebook social graph," *arXiv preprint arXiv:1111.4503*, 2011. 23

[69] D. V. Foster, J. G. Foster, P. Grassberger, and M. Paczuski, "Clustering drives assortativity and community structure in ensembles of networks," *Physical Review E*, vol. 84, p. 066117, Dec. 2011. 23, 146

[70] A. L. Traud, P. J. Mucha, and M. A. Porter, "Social structure of Facebook networks," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 16, pp. 4165–4180, 2012. 23, 146

[71] F. Heider, "Attitudes and cognitive organization.," *The Journal of Psychology*, vol. 21, pp. 107–112, Jan. 1946. 24, 35

[72] "Stanford Network Analysis Project." 25

[73] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proceedings of the 19th International Conference on World Wide Web - WWW '10*, (New York, NY, USA), p. 641, ACM Press, Apr. 2010. 25

[74] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Signed networks in social media," in *Proceedings of the 28th International Conference on Human Factors in Computing Systems - CHI '10*, (New York, NY, USA), p. 1361, ACM Press, Apr. 2010. 25

[75] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, "Dynamical and correlation properties of the Internet," *Physical Review Letters*, vol. 87, p. 258701, Nov. 2001. 28, 41

[76] M. Newman, D. Watts, and S. Strogatz, "Random graphs with arbitrary degree distributions and their applications," *Physical Review E*, vol. 64, p. 026118, 2001. 29, 60

[77] P. M. Blau, "A macrosociological theory of social structure," *American Journal of Sociology*, vol. 83, no. 1, pp. 26–54, 1977. 35, 42

[78] F. Harary, "On the notion of balance of a signed graph," *Michigan Mathematical Journal*, vol. 2, pp. 143–146, 1953. 35, 43, 49

[79] C. Altafini, "Consensus problems on networks with antagonistic interactions," *Automatic Control, IEEE Transactions on Automatic Control*, vol. 58, no. 4, pp. 935–946, 2013. 35

[80] J. Park and M. Newman, "Origin of degree correlations in the Internet and other networks," *Physical Review E*, vol. 68, p. 026112, Aug. 2003. 41

[81] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, "Extracting large-scale knowledge bases from the web," in *VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK* (M. P. Atkinson, M. E. Orlowska, P. Valduriez, S. B. Zdonik, and M. L. Brodie, eds.), pp. 639–650, Morgan Kaufmann, 1999. 41, 43

[82] R. Xulvi-Brunet and I. Sokolov, "Reshuffling scale-free networks: From random to assortative," *Physical Review E*, vol. 70, no. 6, p. 066102, 2004. 41, 45

[83] P. Doreian and A. Mrvar, "A partitioning approach to structural balance," *Social Networks*, vol. 18, pp. 149–168, 1996. 42

[84] G. Facchetti, G. Iacono, and C. Altafini, "Computing global structural balance in large-scale signed social networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 52, pp. 20953–20958, 2011. 42

[85] P. Doreian and D. Krackhardt, "Pre-transitive mechanisms for signed networks," *Journal of Mathematical Sociology*, vol. 25, pp. 43–67, 2001. 42

[86] P. Doreian and A. Mrvar, "Partitioning signed social networks," *Social Networks*, vol. 31, pp. 1–11, 2009. 42

[87] G. Bianconi and A. L. Barabási, "Competition and multiscaling in evolving networks," *Europhysics Letters*, vol. 54, no. 4, pp. 436–442, 2001. 45

[88] J. A. Davis, "Clustering and structural balance in graphs," *Human Relations*, vol. 20, pp. 181–187, 1967. 49, 51

[89] B. Uzzi and J. Spiro, "Collaboration and creativity: The small world problem," *American journal of sociology*, vol. 111, no. 2, pp. 447–504, 2005. 59

[90] T. M. Amabile, "The social psychology of creativity: A componential conceptualization.," *Journal of personality and social psychology*, vol. 45, no. 2, p. 357, 1983. 59

[91] M. D. Mumford, K. S. Hester, and I. C. Robledo, "Creativity in organizations-chapter 1: Importance and approaches," 2012. 59

[92] N. Anderson, K. Potočnik, and J. Zhou, "Innovation and creativity in organizations: A state-of-the-science review, prospective commentary, and guiding framework," *Journal of Management*, vol. 40, no. 5, pp. 1297–1333, 2014. 59

[93] J. Moody, "The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999," *American sociological review*, vol. 69, no. 2, pp. 213–238, 2004. 60

[94] J. E. Perry-Smith, "Social yet creative: The role of social relationships in facilitating individual creativity," *Academy of Management journal*, vol. 49, no. 1, pp. 85–101, 2006. 60

[95] N. Lin, K. S. Cook, and R. S. Burt, *Social capital: Theory and research*. Transaction Publishers, 2001. 60

[96] N. Crossley, S. McAndrew, and P. Widdop, *Social networks and music worlds*, vol. 126. Routledge, 2014. 60

[97] S. McAndrew and M. Everett, "Music as collective invention: A social network analysis of composers," *Cultural Sociology*, vol. 9, no. 1, pp. 56–80, 2015. 60

[98] M. Á. Serrano, M. Boguná, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," *Proceedings of the national academy of sciences*, vol. 106, no. 16, pp. 6483–6488, 2009. 67

[99] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of statistical mechanics: theory and experiment*, vol. 2008, no. 10, p. P10008, 2008. 67

[100] M. De Domenico and A. Arenas, "Researcher incentives: Eu cash goes to the sticky and attractive," *Nature*, vol. 531, no. 7596, pp. 580–580, 2016. 77

[101] L. M. Bettencourt, J. Lobo, D. Strumsky, and G. B. West, "Urban scaling and its deviations: Revealing the structure of wealth, innovation and crime across cities," *PloS one*, vol. 5, no. 11, p. e13541, 2010. 85

[102] B. Klimt and Y. Yang, "Introducing the enron corpus," 2004. 91

[103] N. Eagle, A. S. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proc Natl Acad Sci USA*, vol. 106, 2009. 91

[104] P. O. Larsen and M. Ins, "The rate of growth in scientific publication and the decline in coverage provided by science citation index," *Scientometrics*, vol. 84, 2010. 91

[105] E. A. Leicht, G. Clarkson, K. Shedden, and M. E. Newman, "Large-scale structure of time evolving citation networks," *Eur Phys J B*, vol. 59, 2007. 91

[106] D. J. Solla Price, "Networks of scientific papers," *Science*, vol. 149, 1965. 91, 92

[107] D. J. Solla Price, "A general theory of bibliometric and other cumulative advantage process," *J Am Soc Inf Sci*, vol. 27, 1976. 91, 92

[108] N. Newman, *Networks: an introduction*. New York: Oxford University Press, 2010. 91

[109] "Goldberg sr, anthony h, evans ts (2014) modelling citation networks. preprint. arxiv:1408.2970." 92

[110] C. Calero-Medina and E. C. Noyons, "Combining mapping and citation network analysis for a better understanding of the scientific development: the case of the absorptive capacity field," *J Informetr*, vol. 2, 2008. 92

[111] C. Catalini, N. Lacetera, and A. Oettl, "The incidence and role of negative citations in science," *Proc Natl Acad Sci USA*, vol. 112, 2015. 92

[112] R. Sinatra, P. Deville, M. Szell, D. Wang, and A. L. Barabási, "A century of physics," *Nat Phys*, vol. 11, 2015. 92, 106

[113] J. R. Clough, J. Gollings, T. V. Loach, and T. S. Evans, "Transitive reduction of citation networks," *J Complex Netw*, vol. 3, 2015. 92, 111

[114] "Clough jr, evans ts (2014) what is the dimension of citation space? preprint. arxiv:1408.1274." 92

[115] "https://publish.aps.org/datasets." 92, 93

[116] M. Tumminello, S. Micciché, F. Lillo, J. Piilo, and R. N. Mantegna, "Statistically validated networks in bipartite complex systems," *PLoS ONE*, vol. 6, 2011. 92, 98, 100

[117] M. Tumminello, F. Lillo, J. Piilo, and R. N. Mantegna, "Identification of clusters of investors from their real trading activity in a financial market," *New J Phys*, vol. 14, 2012. 92, 98

[118] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proc Natl Acad Sci USA*, vol. 106, 2009. 92, 102

[119] G. Kossinets and D. J. Watts, "Origins of homophily in an evolving social network," *Am J Sociol*, vol. 115, 2009. 92, 102

[120] P. F. Lazerfeld and R. K. Merton, "Friendship as a social process: a substantive and methodological analysis," 1954. 92, 102

[121] J. Schummer, "Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology," *Scientometrics*, vol. 59, 2004. 94

[122] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J R Stat Soc, Ser B, Methodol*, vol. 57, 1995. 100, 155

[123] A. B. Jaffe and M. Trajtenberg, *Patents, citations, and innovations: a window on the knowledge economy*. Cambridge: MIT Press, 2002. 111

[124] C. Sternitzke, A. Bartkowski, and R. Schramm, "Visualizing patent statistics by means of social network analysis tools," *World Pat Inf*, vol. 30, 2008. 111

[125] J. H. Fowler, T. R. Johnson, J. F. Spriggs, S. Jeon, and P. J. Wahlbeck, "Network analysis and the law: measuring the legal importance of precedents at the us supreme court," *Polit Anal*, vol. 15, 2007. 111

[126] J. H. Fowler and S. Jeon, "The authority of supreme court precedent," *Soc Netw*, vol. 30, 2008. 111

[127] O. Penner, R. K. Pan, A. M. Petersen, K. Kaski, and S. Fortunato, "On the predictability of future impact in science," *arXiv preprint arXiv:1306.0114*, 2013. 114

[128] R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási, "Quantifying the evolution of individual scientific impact," *Science*, vol. 354, no. 6312, p. aaf5239, 2016. 114

[129] B. F. Jones and B. A. Weinberg, "Age dynamics in scientific creativity," *Proceedings of the National Academy of Sciences*, vol. 108, no. 47, pp. 18910–18914, 2011. 114

[130] D. Wang, C. Song, and A.-L. Barabási, "Quantifying long-term scientific impact," *Science*, vol. 342, no. 6154, pp. 127–132, 2013. 114

[131] A. Rzhetsky, J. G. Foster, I. T. Foster, and J. A. Evans, "Choosing experiments to accelerate collective discovery," *Proceedings of the National Academy of Sciences*, vol. 112, no. 47, pp. 14569–14574, 2015. 114

[132] A. Clauset, S. Arbesman, and D. B. Larremore, "Systematic inequality and hierarchy in faculty hiring networks," *Science advances*, vol. 1, no. 1, p. e1400005, 2015. 114

[133] A. Ma, R. J. Mondragón, and V. Latora, "Anatomy of funded research in science," *Proceedings of the National Academy of Sciences*, vol. 112, no. 48, pp. 14760–14765, 2015. 114

[134] T. Jia, D. Wang, and B. K. Szymanski, "Quantifying patterns of research-interest evolution," *Nature Human Behaviour*, vol. 1, p. 0078, 2017. 114

[135] D. Cartwright, F. Harary, and R. Norman, "Structural models: An introduction to the theory of directed graphs," *New York*, 1965. 118

[136] K. Carley, "Extracting culture through textual analysis," *Poetics*, vol. 22, no. 4, pp. 291–312, 1994. 128

[137] R. Axelrod, "The dissemination of culture a model with local convergence and global polarization," *Journal of Conflict Resolution*, vol. 41, no. 2, pp. 203–226, 1997. 136

[138] D. Centola, J. C. Gonzalez-Avella, V. M. Eguiluz, and M. San Miguel, "Homophily, cultural drift, and the co-evolution of cultural groups," *Journal of Conflict Resolution*, vol. 51, no. 6, pp. 905–929, 2007. 136