DISS. ETH Nr. 25138

STRUCTURE AND DYNAMICS OF COLLABORATIVE KNOWLEDGE NETWORKS

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

presented by
VAHAN NANUMYAN

MASTER OF SCIENCE ETH IN PHYSICS ETH ZURICH CH

born on FEBRUARY 7, 1990

citizen of REPUBLIC OF ARMENIA

accepted on the recommendation of
PROF. DR. DR. FRANK SCHWEITZER
PROF. DR. RENAUD LAMBIOTTE
PROF. DR. MATÚŠ MEDO

2018

ETH zürich

Acknowledgments

A Doctoral dissertation is the culmination of a person's long educational journey. Inevitably, the course of this journey is affected by many people. Mine is not an exception, and this is the perfect place to reflect about it.

It all starts in the family. I would not be where I am now, if it had not been the continuing support of my parents in all stages and aspects of my life. Since childhood I enjoyed observing my father work. I found fascinating the synergy between his art and technical know-how, which I try to replicate in my work. My accomplishments are fuelled by my mother's endless energy. She has always been the warrior helping to forge my confidence, the ability to work hard and to face any difficulties in life.

I had great mentors since high school. My first physics teacher, Levon Sanoyan, had an enormous influence towards me becoming a physicist. Then, it was Meruzhan Harutyunyan who cemented this choice. Meruzhan continued to be my advisor during my Bachelor's at Yerevan State University, together with Ashot Gevorgyan, to whom I also owe a huge intellectual debt.

I first came to ETH Zurich to pursue a Master's in physics. Early on I got to know Frank Schweitzer. From attending his lectures, to assisting in teaching, to my Master's projects with him, to finally writing this Doctoral dissertation, I enjoyed learning from Frank how science is done. I would not be half as confident about my academic skills, if I did not work under Frank's strict and high academic standards. I applaud his own academic integrity that goes much beyond the accepted standards and his talent to teach it to others.

I have had exceptional colleagues at the Chair of Systems Design. I can confidently say that I learned something from each of them. Many became my friends. I thank you all for providing the most inspiring and joyful environment in all these years. I am especially grateful to my collaborators, Giona Casiraghi, Antonios Garas, Pavlin Mavrodiev, Ingo Scholtes, Simon Schweighofer and Christian Zingg. To Simon I owe a particular debt as he was always there in the final and most demanding period of writing this dissertation, reviewing it and providing invaluable input. On this note, I must also thank Julia Shumeyko and my wife Alisa for helping to make this dissertation more pleasant to read.

A special gratitude goes to the co-examiners of my dissertation, professors Renaud

Lambiotte and Matúš Medo, who put a tremendous effort to read it with utmost attention and gave invaluable feedback for improving it.

What greatly helped me to stay fully focused on this project for more than three years, were my trainings in Aikido Ikeda Dojo. I am very thankful to all practice partners and the great teachers that made me feel at home in Zurich.

Closing the loop, I want to come back to my loved ones. There are no words to express my gratitude to my parents Lev and Nora and to my sister Mariam, for their unconditional love. I also thank all my friends who enrich my life, Edgar, Gavrush, Vahagn, Ghazar, Aleksan, Narek, Abel, Stefan, Anna, and others whom I am not able to mention now. Last but by no means least, I thank my dear wife and the greatest friend Alisa who continually fills my life with limitless love, happiness and purpose and who has gifted us our first-born Levon only days after I defended this dissertation.

Thank you all!

To my Family

Contents

Abstract

A major means to encode and share scientific knowledge are publications, which cite each other and which are authored by one or more scientists. Citation networks of publications are commonly used to proxy the structure of scientific knowledge. Coauthorship networks are used to represent the social network between collaborating scientists. Yet, these two networks are rarely considered together even though they are interconnected. The multilayer *collaborative knowledge network* that results from combining the two allows us to study how the social relations among authors affect the structure and dynamics of the citation layer. To address this issue, we apply network theory.

In the first part, we analyse the *structure* of collaborative knowledge networks. Our goal is to study dyadic interactions between individual pairs of authors in the context of the whole network. The ability to perform such a study will allow investigating individual citation behaviours of authors, as well as their deviations from community standards. For this, we develop a novel statistical method to extract how much authors' citations to each other deviate from a certain expectation. It builds on three methodological contributions. The first one is a flexible probabilistic model for complex networks that can encode heterogeneity in dyadic interactions. The second one is a procedure to formulate statistical null models for networks that respect temporal ordering of nodes and community structures. The third contribution is a new nonparametric probabilistic measure to quantify the deviation of an observed value from a distribution. With this method at hand, we present the deviations of authors' citations from the expectation formed based on the behaviour of the community at large. We also show how to use these deviations to highlight the intricate sub-community structures within the larger communities.

In the second part, we study the *evolution* of collaborative knowledge networks. We show that the often neglected social layer has a significant effect on the citation layer. Particularly, we find that the overall likelihood of a publication to be cited scales with the number of previous publications by its authors, as well as with the number of their previous collaborators. To obtain this finding, we develop a method to fit and compare probabilistic growth models of multilayer networks. We further look into how the citations are distributed over time for a given publication and we find that citations arrive faster for the authors with more collaborators and more publications.

The scientific contribution of this thesis is twofold. First, we develop novel statistical

methods to study evolving multilayer complex networks. These methods can be applied in various fields. Second, we apply these methods to study citation and collaboration networks from the unified viewpoint of a multilayer network, which leads us to findings that could not be reached by merely considering the two layers in isolation.

Kurzfassung auf Deutsch

Das vorrangige Mittel um wissenschaftliche Erkenntnisse festzuhalten und zu verbreiten sind Publikationen. Publikationen zitieren sich gegenseitig und werden von mindestens einem Autor verfasst. Zitationsnetzwerke wissenschaftlicher Publikationen werden üblicherweise als Abbild der Struktur wissenschaftlicher Erkenntnis verwendet, während Netzwerke von Koautoren als Abbild des sozialen Netzwerks zwischen kollaborierenden Wissenschaftlern gesehen werden. Obwohl diese beiden Netzwerke eng miteinander verbunden sind, werden sie kaum gemeinsam analysiert. Das mehrere Ebenen umfassende kollaborative Wissensnetzwerk, welches sich aus der Kombination der beiden Netzwerke ergibt, erlaubt es uns zu untersuchen, wie soziale Beziehungen zwischen den Autoren die Struktur und Dynamik der Zitier-Ebene beeinflussen. Wir untersuchen diese Fragestellung mit Methoden der Netzwerktheorie.

Im ersten Teil dieser Dissertation analysieren wir die *Struktur* kollaborativer Wissensnetzwerke. Unser Ziel ist es dyadische Interaktionen zwischen einzelnen Forscherpaaren im Kontext des gesamten Netzwerkes zu untersuchen. Die Fähigkeit eine solche Studie durchzuführen wird sowohl die Erforschung des Zitationsverhaltens Einzelner als auch die Abweichung von Standards in Wissenschaftsgemeinschaften ermöglichen. Zu diesem Zweck entwickeln wir eine neuartige statistische Methode, mittels der wir signifikante Abweichungen in der Tendenz von Forschern sich gegenseitig zu zitieren erkennen können. Diese Methode baut auf drei Beiträgen auf. Der erste Beitrag ist ein flexibles probabilistisches Modell komplexer Netzwerke, welches Heterogenität in dyadischen Interaktionen miteinbeziehen kann. Der zweite Beitrag ist ein Verfahren zur Erstellung statistischer Nullmodelle für Netzwerke, welches die zeitliche Reihenfolge der Knoten und Gruppen berücksichtigt. Der dritte Beitrag ist ein non-parametrisches statistisches Mass der Abweichung eines beobachteten Wertes von einer Verteilung. Mithilfe dieser Methoden untersuchen wir die Zitationsmuster zwischen Autoren in verschiedenen Forschungsfeldern. Weiterhin zeigen wir wie diese Abweichungen genutzt werden können um komplexe Strukturen zwischen Untergemeinschaften hervorzuheben.

Im zweiten Teil der Dissertation konzentrieren wir uns aud die *Evolution* kollaborativer Wissensnetzwerke. Wir zeigen dass die oft vernachlässigte soziale Ebene einen signifikanten Einfluss auf die Ebene der Zitationen aufweist. Die Wahrscheinlichkeit einer Publikation zitiert zu werden steigt mit der Anzahl vorhergehender Publikationen des Autors, sowie mit der Anzahl vorheriger Mitautoren. Diese Ent-

deckung basiert auf einer neuartigen Methode zur Kalibrierung und zum Vergleich von probabilistischen Wachstumsmodellen von Mehrebenen-Netzwerken. Weiterhin untersuchen wir die zeitlichen Veränderungen in der Zitationsrate und stellen fest, dass Zitationen sich schneller bei Autoren ansammeln welche über mehr Koautoren und Publikationen verfügen.

Der wissenschaftliche Beitrag dieser Dissertation ist zweifach: Erstens entwickeln wir neuartige statistische Methoden zur Analyse von sich verändernden komplexen Mehrebenen-Netzwerken. Diese Methoden können auch in vielen anderen Bereichen Anwendung finden. Zweitens wenden wir diese Methoden auf die Untersuchung von Zitier- und Mitautor-Netzwerken unter der vereinheitlichenden Perspektive eines Mehrebenen-Netzwerks an. Dieses Vorgehen ermöglicht uns einen Erkenntnisgewinn welcher durch die getrennte Betrachtung der beiden Netzwerke nicht möglich ist.

Chapter 1

Introduction

In order to be effective, policies on how to govern and invest in science must be well-informed. In particular, the mechanisms of how scientific knowledge evolves must be understood. Advancing this knowledge often involves collaborative effort from scientists. In this dissertation, we investigate the co-evolution and interdependence of scientific knowledge and collaborations from the perspective of network theory.

Scientific publications, along with their references to each other, can be considered a representation of scientific knowledge. In these publications, scientists state the results of their research, describe their methods, review previously published literature and discuss the relation to their own research. Sharing the knowledge, obtained from conducting research, is the main goal of publishing scientific papers.

Published papers affect future research, e.g., by acting as inspiration for new ideas. This future research results in new publications. Ideally, the authors of a new publication refer to older publications that inspired them or are otherwise closely related to their own. The formal way to state a reference to a publication is a *citation*. The publications and the citations between them form a directed network, called *citation network*. Readers who use citations to navigate in the large body of scientific literature are effectively traversing the citation network. Because of this, the citation network can be seen as an instance of a *knowledge map* of science [112].

There are different disciplines in science, such as natural and formal sciences, social

sciences, humanities and medicine. Research differs across disciplines in many ways: studied problems, methods, applications, culture, etc. Each discipline is further divided into specific research fields and topics. The body of scientific knowledge can be seen as embedded in an abstract *knowledge space* [112, 167, 191], the dimensions of which correspond to different scientific concepts [194, 220]. The more concepts two research topics have in common, the closer they are in this space.

In practice, scientific publications are often classified according to a semantic scheme [90]. These schemes are an explicit representation of the corresponding knowledge space, in which the publications have specific positions. For example, many journals of physics used for a long time the *Physics and Astronomy Classification Scheme* (PACS), while the *JEL classification codes* are used in economics. The abstract knowledge space can be represented in ways other than predefined schemes. For example, natural language processing allows unsupervised extraction of the main concepts in a publication. Then, the set of all concepts from a corpus of publications—for instance, all of scientific literature—defines the knowledge space. As an alternative to citation networks, concept-based knowledge spaces are commonly used to map science [31, 58, 59, 113].

The interactions of authors within a scientific collaboration can be very intricate [38]. When conducting research, scientists interact in different manners: they discuss ideas, divide and coordinate workload, teach skills by showing, build and operate hardware together, etc. Unfortunately, it is impossible to know how precisely these collaborations unfold, at least at a large scale. Instead, we have to resort to the representation of collaborations by coauthorship relations. These are built from the often long list of authors on a publication.

It is important to note at this point that the content of scientific publications does not represent all of the scientific knowledge but only the *explicit knowledge*. From a widely accepted view introduced by Polanyi, knowledge also has a *tacit dimension* [155] comprising all the aspects that are not possible to encode formally. This kind of knowledge can be shared only through joint practice [53, 108]. A popular elucidating example for the tacit knowledge is riding a bicycle: one can teach another by showing (joined practice), but never by providing, say, a written algorithm of how to keep the balance on the bicycle. One can consider the collaboration relations between scientists as an indication of sharing tacit knowledge. Hence, if the citation network is a map of explicit scientific knowledge, the *coauthorship network*—the proxy for collaborations—can be seen as a map over which tacit knowledge is shared [22].

The above discussion about scientific knowledge can be extended to other domains. A closely related one is Research and Development, where the patents encode the explicit knowledge. Similar to scientific publications, patents usually reflect a collaborative effort of a team of inventors. As in science, there are different sectors in R&D, e.g., pharmaceuticals, computer software and hardware, electronic components, etc. Therefore, the outcomes of science and R&D are prominent examples of *collaborative knowledge spaces*. Throughout the dissertation we interchangeably refer to scientific publications and patents, or, more generally, to *knowledge artifacts*. Similarly, we discuss *authors* of knowledge artifacts, which also refers to inventors in the the case of R&D networks.

1.1 Collaborative knowledge networks

In this dissertation, we investigate collaborative knowledge spaces from the perspective of network theory. Thus, we use the term *collaborative knowledge network*.

Although both citation and coauthorship networks are widely investigated in separation, we argue that studying them together will lead to new insights about collaborative knowledge spaces. The main reason for this is the fact that the dynamics of these two networks are strongly coupled. Over time, new authors enter the coauthorship network when adding new knowledge artifacts into the citation network. Other authors leave the network, and the artifacts contributed by them age, attracting fewer citations [105]. It is also known that the structure of one network may affect the dynamics of the other network. For example, in the context of scholarly publications, previous collaborations among scientists are known to facilitate future citations between them [165].

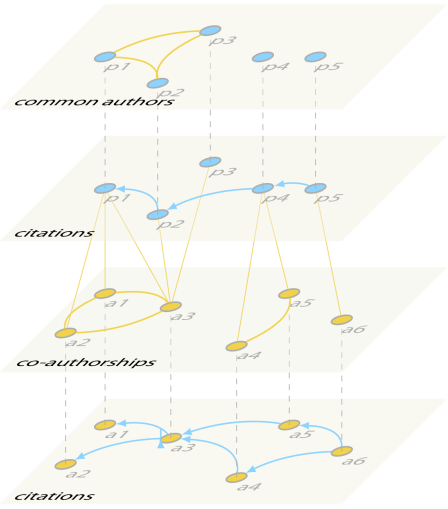


Figure 1.1: Multi-layer network representation of interconnected citation and coauthorship networks. Blue nodes are knowledge artifacts and yellow nodes are authors (inventors).

To study the two networks together, we represent them as a so-called *multi-layer network* [208], shown in Fig. 1.1. Each of the two essential layers—the citation layer and the coauthorship layer—comprises a different type of nodes. In the citations layer (second layer from the top in Fig. 1.1), nodes represent knowledge artifacts (blue in the figure). In the coauthorship layer, nodes represent authors (yellow in the figure). All connections between nodes stem from two base types of edges: the *citation* relations between knowledge artifacts and the *authorship* relations between authors and the artifacts. Edges of the latter type span between two layers. The authorship relations are projected into coauthorship relations between authors (third layer in Fig. 1.1). Additional layers may also be constructed. These can be projections of one layer onto another through inter-layer edges. For instance, a useful projection is the author–author citation network, shown as the bottom layer in Fig. 1.1. Other layers can represent additional relations between nodes, such as *topical similarities* between knowledge artifacts or authors (see Section 5.1.3).

The focus of this dissertation is to investigate how the layers of the collaborative knowledge network evolve together and influence each other. To achieve this, suitable methodologies are developed. Whenever possible, the methodology is developed in a general manner, such that it can be applied or adapted for applications beyond collaborative knowledge networks. For instance, the network ensembles and the corresponding hypothesis testing procedure developed in Chapter 3 are useful for studying any co-occurrence data. Similarly, the new measure of deviation introduced in Chapter 4 can be used to infer signed relations in any data on pairwise interactions. The approach to modelling network growth described in Chapter 6 can be used to study any growing multi-layer networks.

1.2 Generative modelling of complex social systems

The networks representing collaborative knowledge spaces are: (i) *large*, (ii) *decentralized*, meaning that there is mostly no central planning and control, (iii) *open* to new contributors and publication, and (iv) *heterogeneous* both in structure and dynamics. These properties generate *complexity* in structure [18, 139]. The structure and size of both citation and coauthorship networks change over time [138]. In the context of a multi-layer network, the dynamics of one layer influences the dynamics of another layer. These bidirectional dependencies lead to feedback loops and non-

linear dynamics—another known source of complexity [12]. All of the above makes collaborative knowledge spaces instances of *complex systems* and their network representations instances of *complex networks*.

The general trend of increasing availability of data popularised an opinion that theoretical and hypothesis-driven research becomes obsolete. However, data do not speak by themselves and there is a large gap between data and knowledge. Instead, newly available data should be utilised to test more precise scientific hypotheses [154, 198, 219]. This dissertation seeks balance between data-oriented and hypothesis-driven approaches. To achieve this balance, we formalize our hypotheses as *generative probabilistic models* [107] and we apply these to large data sets on scientific publications and patents (see Chapter 2).

In probability and statistics, there are two approaches to finding dependencies between the variables of interest. The first approach is called *discriminative modelling*. Statistical classification and regression fall under this approach. In discriminative modelling, one starts with the *observed data* and aims at identifying the conditional probability to observe the dependent variable, given the independent variable. The second approach is *generative modelling*, which aims at a deeper understanding of the relations between the variables. The objective is to find the joint probability underlying the observed data. Hence, a successful generative model can shed light on the real process generating the data.

Social aspects A publication should be cited when it is of high relevance to the citing one—this and other mechanisms driving the formation of collaborative knowledge networks we have discussed so far can be characterised as *normative aspects* in science [16]. Many quantitative studies in bibliometrics and scientometrics implicitly or explicitly account only for such aspects, e.g., when measuring the scientific impact of a publication by its number of citations [19, 102, 161, 199, 204]. Yet, science is done by scientists who are, like all humans, susceptible to subjective biases. Clearly, citation behaviour can be influenced by a large number of *social aspects*. Below we only mention three aspects, which have been shown to universally influence social relations.

The first of these social influences is *homophily* [122]. Homophily is the universal tendency of human beings to preferentially connect with others who are similar to them, e.g., with respect to ethnicity, language, religion, age, education, occupation, gender, and political position. The influence of homophily can be observed in regard

to a variety of social relations, including marriage, work relations, friendship, and information exchange. The universality of homophily strongly suggests that citation relations between scientists should also be subject to its influence. For our purposes, we have to distinguish between two kinds of homophily. The first one is based on the similarity of the research fields of two scientists. If two scientists work on very similar subjects, they will cite each other more often than scientists who work in very different fields. While still a form of homophily (similarity leading to interactions), this kind of behaviour cannot be attributed to subjective biases. There is no reason why we would expect researchers from different disciplines to cite each other as often as researchers who work on the same subject. In contrast, homophily based on non-scientific criteria, such as race, nationality, or religion, can be called *social bias*.

The second social influence factor is *reciprocity*. Researchers distinguish between positive and negative reciprocity. Positive reciprocity denotes the tendency of an individual to return a positive action, such as a favour, by another individual in kind. However, if an individual feels that her interests were harmed by an action of another individual, she will strive to retaliate—this is what is called negative reciprocity. Sociologists have long identified reciprocity as one of the most important drivers of social interactions [84]. More recently, the importance of reciprocity has also been recognized in game theory and behavioural economics [6, 21]. Repeated acts of positive reciprocity can give rise to friendship relations, whereas repeated acts of negative reciprocity can trigger feuds. The universality of reciprocity in human interactions makes it likely that it also affects science, particularly the citation behaviour of scientists.

While homophily and reciprocity focus on dyadic (pairwise) relations, the third social factor, *structural balance*, concerns triadic relations, i.e., relations between three individuals. It assumes the existence of *positive and negative relations* between individuals. Structural balance postulates that there are unstable and stable triads, with the former having a tendency to transform themselves into the latter over time [34, 50]. A triad is stable either if it consists of three friends, or two friends with a common enemy. Two enemies having a friend in common, as well as three enemies, form an unstable triad. These postulates for social networks build upon a psychological theory that relates a person's cognitive consistency or dissonance to the structural balance of her liking/disliking relationships with various concepts, objects or people [89]. This principle has been known since ancient times, as illustrated by the proverbs "the friend of my friend is my friend", "the enemy of my friend is my enemy", and "the enemy of my enemy is my friend". In social networks, which predominantly represent positive social

relations (such as friendships), structural balance is expressed as *triadic closure*, i.e., the tendency of two individuals sharing a friend to form a friendship of their own [83].

Research gap We have a reason to believe that the three aspects discussed—as well as other social aspects—exert influence on citation behaviour, introducing what we call social biases. The overwhelming majority of studies addressing the social aspects in science are qualitative [16, 47, 103, 158]. For instance, Merton discusses at length the psychological and social mechanisms involved in the reward and communication in science [126]. He himself acknowledges the need for quantitative investigation of the issues he raises. In another example, Larivière *et al.* analyse the homophily in collaborations due to language and geographical proximity, but they do so at a highly aggregated level [106].

To *disentangle* social biases from normative aspects is a big challenge. Let us assume an observation that two scientists cite each others' publications very frequently. We might explain this observation with homophily due to, e.g., the two authors coming from the same country or graduating from the same university. Or we can explain it with reciprocity, i.e., the two authors exchange citations with each other as favours. Finally, it might be a product of triadic closure: the two authors have many friends in common, thus they start to exchange citations with each other, effectively forming what is known as a *citation cartel* [67]. However, it could also be that the high citation count between the two scientists is not an outcome of a social bias, but a result of normatively justifiable aspects. Namely, both authors work on very close subjects, publishing papers highly relevant for each other. Thus, they cite each other frequently for purely scientific reasons. It is not a simple task to identify which of the scenarios is the likely cause of the high observed number of citations. We believe that advanced statistics and generative modelling techniques are necessary in order to distinguish between different aspects leading to the observed citations and coauthorship relations. The following statement summarises the broad *research gap*, which this dissertation attempts to bridge.

There is a lack of quantitative techniques for disentangling normative aspects from social aspects in collaborative knowledge networks.

Network ensembles as null models Let us now briefly illustrate how generative modelling by means of network ensembles can help to address the stated research gap. If we only consider two authors in isolation, we cannot separate the effects of normative and social aspects on their citation behaviour. Instead, we must look into their behaviour towards each other and compare this to the behaviour towards other members of the scientific community.

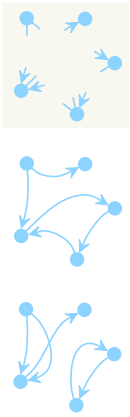


Figure 1.2: Configuration model

If two authors write and cite many publications, we would expect them to cite each other more frequently by mere chance than two less prolific authors. Such difference in expectation is due to *combinatorial effects*. This expectation can be quantified using a *statistical network ensemble*, which is the collection of all possible networks satisfying certain conditions. We will provide the formal definition of network ensembles in Section 2.4.1. A network that is part of the ensemble is often referred to as a *realisation of the ensemble*. The ensemble also describes the probability of each realisation. To familiarize the reader with the concept, let us introduce one such ensemble called *configuration model* [129]. All networks in this ensemble have the same total number of nodes and edges. Moreover, each node has a fixed number of edges across all networks. What changes from network to network within the ensemble, is to which other nodes these edges are connected. Figure 1.2 illustrates a configuration model for five nodes and five edges. The ensemble sets the number of outgoing and incoming edges for each node. The top panel in Fig. 1.2 presents these in the form of severed edges, or "stubs". Then, the realisations of the ensemble are formed by randomly connecting the outgoing stubs to incoming stubs. The bottom two panels in Fig. 1.2 show two such network realisations. The configuration model does not put any additional constraints, so all the networks that are part of the ensemble are equally probable.

Now, how do we use a network ensemble to differentiate between the aspects influencing the formation of collaborative knowledge networks? The outlined configuration model can already account for the combinatorial effects. Specifically, an ensemble with fixed number of citations that each author gives and receives describes all possible ways that these citations can be distributed among the authors. If we focus on a certain pair of nodes (authors), the ensemble provides the distribution of the citation counts between the two authors. This is illustrated in Fig. 1.3 for a network of five authors, where the thickness of the edges is proportional to the number of corresponding

citations. According to the shown distribution, the most probable number of citations from author 2 to author 3 is equal to six. Then, we can colour the values in the distribution according to how far they are from a central value (e.g., expectation, median or mode). If the value is probabilistically much larger than the central value, we observe *significant over-citation* (shown in green). Similarly, we observe *significant under-citation* (shown in red) if the value is much smaller than the expectation. Under-citations have been used to address the problem of identifying *missing links* [42]. According to the illustration, the observed ten citations are only non-significantly more than expected. The deviation of observed number of citations against the distribution can be described in terms of a *signed relation* between two authors. A positive sign corresponds to over-citation and a negative sign to under-citation. The absolute value describes the extent of the deviation.

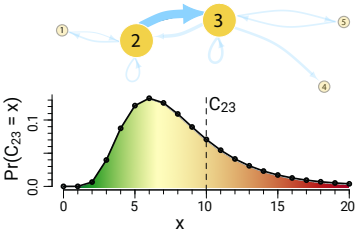


Figure 1.3: Illustration for the probability distribution of the number of citations from author 2 to author 3 according to a network ensemble.

When a network ensemble is used like in the procedure above—that is, to compare the observed network to other networks with matching certain characteristics—it is said to be a *null model*. The differences found between the observed network and the ensemble in such a comparison highlight the characteristics of the observed network that cannot be explained by the conditions that define the ensemble. In other words, these conditions are "nullified" in the comparison. In our comparison of the empirically observed number of citations to the expected from the configuration model, we nullify the combinatorial effect of the total number of edges each node has. That is, if we find that the observed number of citations between a given pair of authors is close to the expectation according to the configuration model, we can state that such citation behaviour can be *expected by mere chance*.

While being good for identifying the combinatorial effects, the comparison of the observed network to the configuration model does not enable us to differentiate between normative and social aspects discussed above. To achieve this ability, we need more sophisticated null models. Specifically, we need a network ensemble that encodes all aspects that we consider normative. Then, with the normative aspects nullified, the deviations from this null model will correspond to social biases. The development of such ensembles is one of the research questions addressed in this dissertation.

1.3 Research questions

There is a large body of literature treating both citation networks [139] and the coauthorship networks [98, 138] as complex networks. There is also a growing number of studies on multi-layer networks [17, 143, 208]. These two lines of research have an unexplored potential: the co-evolution of citation and coauthorship networks can be studied from the multi-layer network perspective.

Network structure In the first part of the dissertation, we study the effect that the positions of authors in the coauthorship network have on the *structure* of the citation network. We have already discussed one scenario of using network ensembles. There are other scenarios for choosing statistical ensembles of random networks to investigate complex systems of many interacting elements. For instance, individual interactions may be unknown, while the macroscopic properties of the system (and the corresponding network) are known. Or, the system may be too large to represent all the interactions to the fullest detail, so a network ensemble is used to compress the information with the "random" part accounting for the discarded details. But most importantly, network ensembles are a perfect choice when the observations of edges include noise—either due to the data collection technique, or due to the intrinsically stochastic nature of the system that the network represents.

A general feature of existing network ensembles is to not distinguish between nodes of a network [20, 136]. That means, the network does no change if two nodes with the same degree are swapped. This assumption of indistinguishable nodes is too strong when representing social systems. For instance, social agents (represented by network nodes) may have intrinsic and unobservable preferences with whom to interact. To the best of our knowledge, there are no analytically tractable network ensembles that allow for heterogeneous preferences for nodes with whom to interact. As discussed earlier, authors in collaborative knowledge networks tend to cite authors who work on similar topics more often, meaning that the topical similarity defines the heterogeneity in citations between different authors. Hence, a new network ensemble is needed to study collaborative knowledge networks. This leads us to the first research question in this dissertation.

RQ 1. Develop a network ensemble for collaborative knowledge networks. The ensemble must account for known explicit heterogeneity in the network structure.

The ensemble proposed in Chapter 3 goes beyond the assumption of indistinguishable nodes. It introduces the notion of *edge propensity*, which expresses the relative preference of one node to connect with another. Because of the dyadic nature of the edge propensity, one can construct a network from it and consider it as a layer in a multi-layer network. And vice versa, different relations between nodes, as well as combinations of these relations, can be used as propensities [35, 36]. In the context of collaborative knowledge networks, we use propensities to encode topical similarity between authors into a network model of citations among authors.

We formulate our network ensemble such that the following question can also be answered.

RQ 1(a). How can we identify significant edges in a network using the network ensemble from *RQ 1*?

To answer this question, we start with the network ensemble that incorporates edge propensities. However, in this case we do not encode prior information (such as topical similarity) into propensities but use these as free parameters, which allows us to fit the ensemble to the observed data. Then, the fitted propensities that are valued above a certain threshold provide us with a so-called *backbone* of the network—a subset of the network comprising particularly strong interactions, which ideally is subject to less noise than the whole network. This approach of inferring the backbone is applicable whenever the data captures repeated interactions between nodes [41, 55].

It is generally a challenge to find a ground truth for *statistical inference problems* (procedures of consecutive model selection, parameter fitting and interpretation). For that reason we employ not only empirical, but also synthetic data sets. For the latter, we develop an agent-based model inspired by location sharing social networks [41]. We then compare the inferred backbone of the co-location network with the ground truth that stems from the rules of the agent-based model.

The two research questions *RQ 1* and *RQ 1(a)* are complementary. In the first question, we use propensities to encode the known heterogeneity in interactions, while in the second, we infer the unknown heterogeneity.

Deviations in network structure We conjecture that the number of citations between two authors bears the footprint of the social biases one author has in favour

of or against the other author. As the first step towards identifying these biases, we assume that a normative expectation can be formulated for how an author should distribute citations to other authors [42]. Once this expectation is formed, the social biases between two authors will be reflected in the deviation of the observed number of citations from the expectation. In a recent study, Ciotti *et al.* make the first step in quantifying the deviations in the network structure from a null model by aiming to identify missing links in the network [42]. Their argument is based on the theory of homophily [123]. To our knowledge, the study was the first to infer *signed relations* [51, 88, 111] from data on unsigned interaction counts. However, the method is tailored for the specific system and the problem of the study and is not easily generalisable. Instead, we develop a general statistical method for inferring signed relations from repeated unsigned interactions. We then provide a procedure to identify significant over- and under-citations among authors. As discussed in the previous section, if all normative factors are accounted for in the null model, these over- and under-citations represent the social biases between authors. We also hypothesise that more prominent authors are more prone to social biases among each for such reasons as, e.g., stronger competition for attention and funding. The prominence of an author will be reflected by a more central position in the network. We pose the problem of identifying social biases and their relation to author centrality as our next research question.

RQ 2. Quantify biases authors have against or in favour of each other. Study the relation between these biases and centralities of authors.

As already discussed, authors are not expected to distribute citations uniformly among each other. Instead, one is expected to cite more frequently an author whose work is topically closer. The network ensemble developed under *RQ 1* is used to find the expected number of citations between each pair of authors. Topical similarities between authors are accounted for by means of edge propensities.

Quantifying the topical similarities between two authors is a challenge for two reasons. First, as mentioned earlier, there is no one absolute way to represent the knowledge space on which the topical similarities are measured. Second, we need an asymmetric measure of similarity. This need is due to the fact that the citations are directed, so two authors can be expected to cite each other differently and the edge propensities must account for this. For the first challenge, we use topological measures of topical similarity, such as *bibliographic coupling* and *co-citations* [212]. To address the second

challenge, we modify an established metric, the *Jaccard index* [95].

Once the network ensemble is constructed, the empirical network is to be compared against it. The comparison shall identify the significant *over-* and *under-citations* between authors. For this, we develop a new method to measure the deviation of an observation against a distribution. For the citation network between authors, the method leads to *signed relations* between them. The absolute value of a positive and negative relation shows the significance of the over- or under-citation, respectively.

Network growth The majority of works on the dynamics of multi-layer networks is related to a special class of *multiplex networks*. Each layer of these networks has a different type of edges but one and the same set of nodes [131, 143]. In the second part of the dissertation, we develop a growth model for multi-layer networks with *different* sets of nodes on different layers and we apply it to investigate the coupled dynamics of coauthorship and citation layers. For collaborative knowledge networks, assuming only the growth—i.e., only the addition and not the removal of nodes and edges—reflects the fact that once a publication is made, in principle, it stays accessible for the scientific community. Technically, *retracted* publications are removed from the network but such retraction are rare, so they do not influence the outcomes of statistical analyses [65].

We further address the issue of attention in science from the perspective of coupled dynamics of the network layers. A recent study by Parolo *et al.* sets a good ground for the this issue by showing that attention, measured in terms of citation rate of an artifact, decays over time [149]. They also quantify general trends of attention decay over a long time period.

There are multiple established and thoroughly studied growth models for networks that describe the growth of citation networks to a different extent [98]. Similarly, models of social network formation [183] and team formation [86] have been proposed and used to describe coauthorship formation. However, all these models do not account for the interdependence of collaboration and citation structures. This leads us to the next research question of this dissertation.

RQ 3. Define a growth model to describe the coupled growth of the collaborative knowledge networks. How does this model compare to modelling the citation and coauthorship networks separately?

The aim is to develop the growth model with a *focus on the coupling* between network layers. To achieve this, established models for each of the two layers are chosen as components in the coupled model. In this way, the model becomes a framework and allows selection and reassessment of one-layer models in the context of coupled growth of a multilayer network. The examples analysed in this dissertation concentrate on the growth of the citation network, dependent on the earlier states of both citation and coauthorship networks. This complements the approach from my Master's thesis [133], where the coauthorship network growth was studied in a similar context. Nevertheless, we provide a general recipe for formulating and analysing the simultaneous growth of the network layers. Additionally, we develop a model selection procedure for the coupled growth. We assess the goodness of a model based on the likelihood of each edge that is added to the network. This is different from the common approach of estimating the model based on aggregate features of the network, such as the final degree distribution [133, 176]. Our model selection procedure has two additional advantages. First, it allows to estimate the error in model parameters—uncommon in modelling network growth. Second, it is scalable to large networks under certain conditions.

Social position of authors and attention decay One aspect of an author's social position in the collaborative knowledge network is measured by centrality in the coauthorship layer. The simplest among these measures is *degree centrality*, defined as the number of edges a node has. In a coauthorship network, degree centrality captures the total number of coauthors of a given author. In scientific communities, the number of coauthors follows a broad distribution [138]. That means that the majority of authors are *peripheral* around a small fraction of highly central authors.

Existing findings in science may trigger new ideas and new research. Most of the times these triggers come from more recent publications. Authors try to be informed about the newest scientific advancements in their field in search of inspiration and to properly position their research. However, it becomes harder and harder for an individual to stay informed due to the accelerating pace of knowledge growth (e.g, measured by the rate of new publications). As a result, it becomes easy to overlook important publications [42].

There is evidence that authors rely on their coauthorship networks in the search of relevant publications. It is shown that authors cite their previous collaborators significantly more often than expected [165]. This leads us to the last research question.

RQ 4. How does the position of an author in the coauthorship network influence the attention towards her knowledge artifacts?

We measure the *attention* towards a publication by the citation rate, i.e., the number of new citations per time unit. On average, the highest citation rate comes a short time after publication and rapidly decays afterwards [149]. We test the hypothesis that authors who have more coauthors get attention towards their new publications faster.

The answers to the research questions posed above can be summarized as a twofold contribution. First, new methods in network theory, which can be applied in various research fields, are introduced. Second, with the help of these methods, the interplay between the social position of authors and the citations to their publications is studied from multiple angles.

1.4 Dissertation overview

Figure 1.4 presents possible paths that the reader can follow in this dissertation. In all cases it is advisable to start from Chapter 2, which introduces the main concepts and the notation, as well as the data sets used in the later chapters. Chapters 3, 4 and 6 assume basic knowledge of probability and statistics and will be of interest for readers seeking new methods of network theory. Chapters 3 and 4 lay the ground for Chapter 5. These three chapters are grouped in one part that addresses the topological properties of networks. Chapter 6 and Chapter 7 address two different aspects of network dynamics. Below is a more detailed description of each chapter.

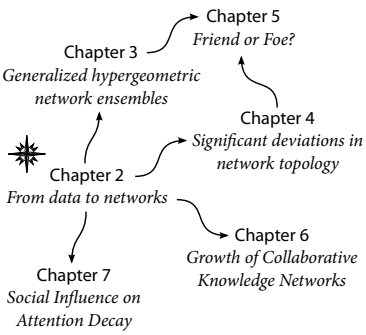


Figure 1.4: Roadmap for reading

Chapter 2 formalises the multi-layer network representation of collaborative knowledge networks. It introduces the notation, as well as the general methods used throughout the dissertation. It also describes the data sets that are studied.

Chapter 3 develops *generalised hypergeometric network ensembles*. The ensemble allows for testing hypotheses about the heterogeneity in edge formation. It introduces a new procedure to extract the backbone of a network using the ensemble.

Chapter 4 presents a novel non-parametric measure of deviation of an observation from a distribution. This measure can be applied to a broad range of probability distributions—continuous or discrete, skewed and/or bounded. The application of this measure to networks is also discussed.

Chapter 5 combines the methods developed in Chapters 3 and 4 to study the structure of collaborative knowledge networks. Specifically, the citation network among authors is investigated. A special formulation of the ensemble is tailored to respect the temporal order of the publications. Multiple definitions of topical similarity between authors are discussed. Finally, over- and under-citations among authors are quantified by what we call a *friend-or-foe matrix*.

Chapter 6 introduces a procedure for analysing mechanisms of the coupled growth collaborative knowledge networks. The procedure relies on *maximum likelihood estimation* to find the parameters of the growth models. The numerical algorithm that provides parameter errors is discussed. The conditions and performance of scaling to large networks are also investigated.

Chapter 7 focuses on the decay of attention towards publications. Linear regression analysis highlights a significant relation between the parameters of the decay and the position of the authors in the collaborative knowledge network.

Chapter 2

From data to networks

Summary

In this chapter we formally introduce the collaborative knowledge networks. We begin with the general mathematical notation of networks and network measures. We discuss the procedure that translates the available data into networks and the caveats of such network representation. In particular, we discuss the meaning of co-authorship edges in the case of very large collaborations. We further describe the three main data sets on patents and scientific publications, which we use throughout the dissertation. Finally, we introduce the methods from statistics that are used in the later chapters.

This chapter has been written specifically for this dissertation.

Knowledge spaces have been represented as networks since the establishment of the fields of *bibliometrics*, *informetrics* and *scientometrics* [159] (these fields have lately been called *science of science* [70]). With the emergence of the field of complex networks over the last three decades, the interest towards network representation of knowledge spaces has grown considerably [160]. With the aim of understanding how science evolves and organizes itself, various methods—from visualisation to generative modelling techniques—have been applied to networks of citations, co-authorships, co-citations, bibliometric coupling, keyword co-occurrence, [57, 70, 159] etc. These networks, collectively known as *bibliometric networks*, are derived from a small set of notions that describe knowledge artifacts (scientific publications, patents). These notions are called *metadata*. The most used elements of metadata to construct the networks are: a unique identifier (e.g., DOI) of the artifact, the list of authors with their affiliations, a list of keywords, publication venue, publication date, and the list of cited publications.

Below we discuss the general procedure of constructing networks from the metadata on knowledge artifacts. But first, we introduce the necessary concepts and notation from network theory.

2.1 Network theory

There is no lack of excellent textbooks and review papers that offer an overview of network theory. In the following, we mostly adopt the notation from the seminal book by Newman [136]. More formally networks are called *graphs*. Correspondingly, network theory is known as *graph theory* in mathematics. A network is an object comprising a set of *nodes*, which are connected by a set of *edges*. Nodes are alternatively referred to as *vertices* and edges as *links*.

A *simple graph* $G = (V, E)$ comprises a set of nodes V and a set of edges $E \subseteq V \times V$. Two nodes i and j are called *neighbours* if they are connected by an edge. The corresponding edge is fully identified by the pair of nodes i and j that it connects, thus we denote it as (i, j). The size of a network is characterised by the number of nodes $n = |V|$ and the number of edges $m = |E|$.

A graph can be described in terms of the *adjacency matrix* $\mathbf{A}$. For a simple graph, the binary entries of the adjacency matrix show if two nodes are connected by an edge or not. That is, $A_{ij} = 1$ if nodes i and j are connected and $A_{ij} = 0$ otherwise. For a simple

graph $\mathbf{A}$ is symmetric, $A_{ij} = A_{ji}$, and has zero entries on the diagonal, $A_{ii} = 0$.

2.1.1 Networks types

Building on the notation for simple graphs, let us now generalise it to networks that are: directed, with self-loops, weighted, signed, multi-edged, time-stamped, bipartite, multiplex and multi-layer.

In a *directed* graph the pair of nodes defining an edge is ordered, meaning that generally $A_{ij} \neq A_{ji}$. A network is said to have *self-loops* if there are edges that connect a node with itself, i.e., $\exists i$ such that $A_{ii} \neq 0$. In a *weighted network*, edges are characterised by an additional attribute, the weight $w : E \to \mathbb{R}^+$. The adjacency matrix of a weighted network incorporates these weights, such that $A_{ij} = w((i, j))$ for $(i, j) \in E$. The network is called signed if the weights are allowed to also take negative values.

An important type of networks are the *multi-edged* networks, or *multi-graphs*. These are the networks that allow multiple parallel edges between a given pair of nodes. The adjacency matrix in this case is integer-valued and represents the multiplicity of edges between the nodes, $A_{ij} \in \mathbb{Z}$. It is important to distinguish weighted networks and multi-graphs, even though both may have the same adjacency matrix representation. The weight is an attribute of the *one and only* edge between a given pair of nodes. In principle, the two concepts of multi-edges and weights are not mutually exclusive, so a multi-graph can be weighted, with each of the parallel edges having its own weight. Other attributes, such as *time-stamps*, can also make parallel edges distinguishable. This said, drawing the multiplicity of edges in multi-edge networks as weights can improve the readability of network visualisations.

A *temporal network*, also known as *dynamic graph* or *time-stamped graph*, is a graph $G^T = (V, E^T)$ that has time-stamped edges $E^T \subseteq V \times V \times [0, T]$ [39, 62, 172]. In many applications, the time-stamped edges are considered to be instantaneous, i.e., present only at one moment in time [172]. Instead, in the case of *growing networks*, the time-stamp corresponds to the moment when the edge is added to the network. In growing networks, nodes are also added over time. Thus, we denote a growing network as $G(T) = (V(T), E(T))$ with $V(T)$ and $E(T)$ being the set of nodes and the set of edges that have been added to the network up to the time T. In contrast with temporal networks, non-time-stamped networks are often called *static*.

A *bipartite network* $G = (V^s, V^t, E^{st})$ represents relations $E^{st} = V^s \times V^t$ between two sets

of nodes V^s and V^t. The matrix representation of a bipartite network is a rectangular *incidence matrix* $\mathbf{B}$ of the shape $|V^s| \times |V^t|$.

It is a common technique to project a bipartite network onto one of the node sets [217]. Let us consider such *one-mode projection* onto the set V^s. Then, two nodes $i, j \in V^s$ are connected with a projected edge if there is a node $k \in V^t$ that they both are connected to in the original bipartite network. Hence, the adjacency matrix A^s of this projection writes as

$$A^s_{ij} = \sum_{k \in V^t} B_{ik} B_{jk}. \tag{2.1}$$

The value of A^s_{ij} can be interpreted both as a weight and a multi-edge. We opt for the latter because, as mentioned above, it can preserve more information. For instance, each of the projected parallel edges can store the corresponding node $k \in V^t$ from which it was projected as an attribute. One-mode projections of a bipartite network create structures known as *cliques*. A clique is such a subset of nodes that any two nodes in it are connected by an edge. In a projection on to V^s there is a clique corresponding to each $k \in V^t$ that comprises all neighbours of k. Because of this property, one-mode projections are instances of *hypergraphs*—objects in which edges connect more than two nodes.

A *multiplex network* $G = (V; E^1, \dots, E^L)$ is a special case of *multi-layer networks* that comprises L layers, each of which hosts one set of edges $E^l \subseteq V \times V$ between one and the same set of nodes.

More generally, a *multi-layer network* is a type of network that comprises multiple types of nodes and/or edges. A multi-layer network $M = (\mathbf{V}, \mathbf{E}, \mathbf{E}^B)$ with L layers, each comprising one set of nodes from $\mathbf{V} = \{V^1, \dots V^L\}$ and one set of edges from $\mathbf{E} = \{E^1, \dots E^L\}$. It also comprises *inter-layer* edges that connect nodes across layers, $\mathbf{E}^B = \{E^{st} \subseteq V^s \times V^t \mid s, t \in \{1, \dots, L\}; s \neq t\}$.

A multi-layer network can be decomposed into separate graphs for each layer, $G^l = (V^l, E^l)$ for $l \in \{1, \dots, L\}$, and bipartite networks between layers, $G^{st} = (V^s, V^t, E^{st})$ for s, t such that $E^{st} \in \mathbf{E}^B$. The set of nodes on some of the layers can be the same, in which case these layers form a multiplex. In that case, the inter-layer links provide a one-to-one correspondence between the nodes on different the layers.

A multi-layer network is described by a *supra-adjacency* matrix that has a block structure [208]. Each block $\mathbf{A}^l$, $l \in \{1, \dots, L\}$, on the diagonal of this matrix corresponds to the adjacency matrix of the layer l in the multi-layer network. Each off-diagonal

block $\mathbf{B}^{st}$ corresponds to the inter-layer connections between two layers s and t. When the two layers have different sets of nodes, the matrix $\mathbf{B}^{st}$ is the incidence matrix of the corresponding bipartite network. When the layers s and t form a multiplex, the corresponding matrix $\mathbf{B}^{st}$ is an identity matrix, i.e., $B^{st}_{ii} = 1$ and $B^{st}_{ij} = 0$ for $i \neq j$.

Different projections can be made in a multi-layer network, similar to one-mode projections of a bipartite network. One way is to project the inter-layer edges onto a one of the layers, in the exact manner as for bipartite networks. The co-authorships layer in Fig. 1.1 on Fig. 1.1 is an example of this. Another way is to project edges on one layer onto the nodes of the other layer through the inter-layer edges. In this case, the projection of an edge $(i, j) \in E^l$ through the inter-layer edges E^{st} is the set of edges $\{(p, q)\}$ on a new layer l' for all $(p, i) \in E^{st}$ and $(q, j) \in E^{st}$. For example, the bottom layer of Fig. 1.1 is the projection of the second layer on the third layer.

The last type of graphs utilised in this dissertation is called *directed acyclic graphs (DAGs)*. A directed acyclic graph is a directed graph $G^{\text{DAG}} = (V, E)$ in which there are no *cycles*—i.e., no sequence of *adjacent edges* originating and ending with the same node $(i, j), (j, k), \ldots, (q, i)$ for $i, j, k, q \in V$. A defining feature of DAGs is the *topological ordering*, which means that there is an order of vertices $v_1, v_2, \ldots v_n$ such that for any $(v_i, v_j) \in E$ the inequality $i > j$ holds. This means that the adjacency matrix of a topologically ordered DAG is a triangular matrix. Some growing networks form DAGs. This happens when the network grows by addition of nodes and when the edges are formed only by the newly added node, pointing in the same direction.

2.1.2 Network measures and metrics

We have already mentioned the simplest measures that characterise the size of a network—the number of nodes n and the number of edges m. These can be generalised for a multi-layer network $M = (\mathbf{V}, \mathbf{E}, \mathbf{E}^B)$ as $\mathbf{n} = \{n^1, \ldots, n^L\}$, $\mathbf{m} = \{m^1, \ldots, m^L\}$ and $\mathbf{m}^I = \{m^{st}, \ldots\}$, where $n^l = |V^l|$, $m^l = |E^l|$ and $m^{st} = |E^{st}|$ for $E^{st} \in \mathbf{E}^B$.

Node degree The number of edges of a node is called *degree*. For a simple graph, i.e., an undirected and unweighted network, the degree of node i is calculated from the adjacency matrix as

$$k_i = \sum_{j=1}^{n} A_{ij}. \tag{2.2}$$

For a directed network, the degree is split into *out-degree* and *in-degree*, corresponding respectively to the number of edges that point away from and towards the node.

$$k_i^{\text{out}} = \sum_{j=1}^{n} A_{ij}, \qquad k_i^{\text{in}} = \sum_{j=1}^{n} A_{ji}. \tag{2.3}$$

The degrees of all nodes add up to twice the number of edges in the case of undirected networks. In the case of directed networks, the sum of all out-degrees equals the sum of all in-degrees, which is equal to the number of edges in the network,

$$m = \sum_{i=1}^{n} k_i^{\text{out}} = \sum_{i=1}^{n} k_i^{\text{in}}. \tag{2.4}$$

The *degree sequence* is a vector of size n comprising the degrees of all nodes in the network, $K^{\text{out/in}}(G) = \{k_1^{\text{out/in}}, \dots, k_n^{\text{out/in}}\}$.

We can generalise the concept of node degree to multi-layer networks. Nodes of each layer have a degree with respect to the edges within the layer and degrees with respect to inter-layers edges. Thus, the node i on layer l will have a degree $k_i^l = \sum_{j \in V^l} A_{ij}^l$ within the layer and degrees $k_i^{lt} = \sum_{j \in V^t} A_{ij}^{lt}$ for each layer $t \neq l$. If some of the layers are directed, the corresponding degrees are substituted by in- and out-degrees, as Eqs. (2.2) and (2.3) show for a single-layer network.

Paths A path on a network is a sequence of nodes in which each consecutive pair is connected by an edge.

$$\pi_{\lambda,ij} = \left\{ v_{h_1}, \dots, v_{h_{\lambda+1}} \right\}, \tag{2.5}$$

where $h_l \in \{1, \dots n\}$ represents the index of l-th node on the path, $h_1 = i$, $h_{\lambda+1} = j$ and $(v_{h_l}, v_{h_{l+1}}) \in E$ for $l \in \{1, \lambda\}$. The *length* of a path, λ, is the number of edges it traverses on the network. In this dissertation, we consider only *self-avoiding* paths, meaning that each node on the path can be visited only once, i.e., $v_{h_l} \neq v_{h_k}$ if $h_l \neq h_k$. For an unweighted network, the number of paths of length λ between a pair of nodes is given by the λ-th power of the adjacency matrix,

$$|\{\pi_{\lambda,ij}\}| = [\mathbf{A}^{\lambda}]_{ij} \tag{2.6}$$

The definition of paths can be extended to multi-layer networks. We are particularly

interested in two cases. The first case comprises the paths of length two that go through inter-layer edges between two layers. They originate and end on the same layer. The second case comprises the paths of length three that originate and end on the same layer but also traverse an edge on the second layer. These two types of paths are related to network projection discussed earlier. Indeed, by comparing Eq. (2.1) and Eq. (2.6) we see that the number of multi-edges between a pair of nodes in the one-mode projection equals to the number of paths of length two in the corresponding inter-layer edges (which form a bipartite network) between the same pair of nodes. Similarly, the paths of length three in the second case correspond to the projection of the edges of one layer onto the nodes of another layer. We will utilize this correspondence throughout the dissertation.

Centrality When studying networks, a common question to answer is how important certain nodes and edges are. Such a measure of importance is called *centrality*. Depending on the context, there are different ways to define node centrality. The simplest one is the *degree centrality*, which is equal to the degree of the node. While the degree centrality captures only the direct neighbourhood of a node, other measures capture more information about the topology of the network. Spectral measures, such as *eigenvector centrality* and *PageRank*, assign higher centrality to nodes that are connected to other central nodes [146]. Other measures are based on paths that pass through a given node. *Closeness centrality* is inversely proportional to the average length of shortest paths from a given node to all other nodes. *Betweenness centrality* captures how many shortest paths between any two nodes pass through a given node. Throughout this dissertation we will use the degree centrality due to its simplicity. However, most of the discussion can be refined by substituting the degree by a different centrality measure.

Node similarity Another important question in network theory addresses how similar nodes are. One approach to measure similarity between nodes is called *structural equivalence*. According to it, the more common neighbours two nodes have, the more similar they are. Two often-used formulations are the *cosine similarity* and the *Jaccard index*. For two nodes, both measures capture the number of common neighbours relative to the total number of neighbours of the two. We use Jaccard index, which is generally defined for two mathematical sets as the number of common elements divided by the number of all unique elements in the two sets. For a simple graph, it

writes as

$$\sigma_{ij} = \frac{\sum_k A_{ik}A_{jk}}{k_i + k_j - \sum_k A_{ik}A_{jk}}. \tag{2.7}$$

To count the number of unique neighbours of the two nodes in the denominator of Eq. (2.7), we corrected for the fact that the sum of two degrees counts the common neighbours twice. The above definition can be extended for directed, as well as for multi-layer networks. In our application in Chapter 5, we will also extend the notion of similarity by making it asymmetric.

The list of network measures and metrics introduced above is not extensive. Only the ones that are relevant for this dissertation are described. Some other measures, such as *reciprocity* and *structural balance* will be introduced later within the corresponding context.

2.2 Building the networks

Following the common approach in scientometrics [24, 57, 59, 112, 191], we use the metadata on knowledge artifacts to study the evolution and structure of collaborative knowledge spaces. In this section, we provide explanation of how we construct collaborative knowledge networks from a collection of knowledge artifacts. Figure 2.1 shows the progression from the original form of the knowledge artifact intended for human readers to the *data model* behind the networks that we construct.

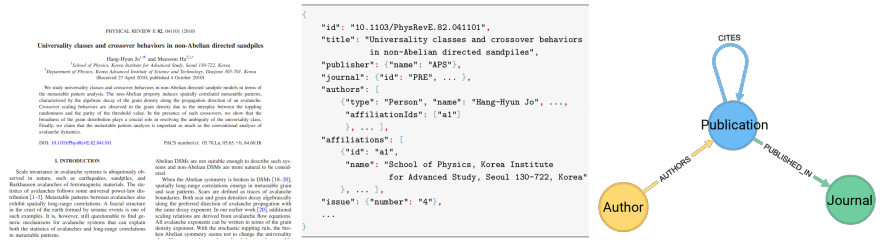


Figure 2.1: (Left) A scientific publication intended for human readers, (middle) its meta-data in machine-readable JSON format, and (right) the data model underlying our networks.

We start our analysis from the metadata of the artifacts. There are various sources providing metadata on large number of knowledge artifacts in a convenient machine-readable format (such as JSON objects, shown in the middle of Fig. 2.1). However,

most of these sources suffer a crucial drawback. In these, the authors of knowledge artifacts are listed by their names and affiliations, which does not guarantee a unique identification of a person across all her publications. Resolving such ambiguity is a large research topic in itself known as *record linkage* and *entity resolution* [13, 114, 210]. Instead of performing the author disambiguation ourselves, we choose data sources that already include the disambiguation. We will discuss the data sources in more detail later in this chapter but we list these here:

- the U.S. patent inventor database for the years 1975–2010 compiled by Li *et al.* [114]
- INSPIRE High Energy Physics (HEP) information system[1]
- physics publications in the journals published by American Physical Society (APS)[2] with author disambiguation provided by Sinatra *et al.* [181]

Only a limited part of the metadata is needed for building our networks. We use a unique identifier for each knowledge artifact. These are digital object identifiers (DOIs) for scientific publications and the United States Patent and Trademark Office (USPTO) patent numbers for the U.S. patents. For each artifact, the metadata provides the list of unique identifiers of the cited artifacts. A unique identifier for each author (inventor) is also provided. In the INSPIRE data, these identifiers are curated by the community using the INSPIRE platform. In the other two data sets, the author names are disambiguated using entity resolution techniques.

Next, we include the journal in which a scientific article is published and the classes under which the patents are categorised. We do so because we want to perform our analyses on networks corresponding to certain scientific fields and patent classes. We will focus on the citations networks within a single journal or patent class for two reasons: (i) to have conveniently small networks for which the analysis is very fast, (ii) to make sure there are no large differences in the community practices whithin one network, the mixture of those could conceal relevant patterns from the analysis. However, we believe that all the analyses and most of the outcomes will hold also for larger networks, e.g., the ones corresponding to a whole data set. We also account for the time-stamps of the knowledge artifacts—the publication dates of scientific publications and the granting date of patents. The time-stamp of a publication is also

[1] The data is released under CC0 license and is available at `https://inspirehep.net`
[2] The data is provided by APS for research, `https://journals.aps.org/datasets`

attributed to all edges attached to it, with the citation edge having the time-stamp of the citing publication. This is necessary for analysing the growth of the networks.

Combining all of the above, we arrive at the data model shown on the right of Fig. 2.1. A citation originates from a publication (blue node and edge) and arrives at a publication; an author authors (yellow node and edge) a publication; and a publication is published in a journal (green node and edge).

Multi-layer collaborative knowledge network The data model described above serves as a blueprint for our collaborative knowledge networks. First, we choose the journal (patent class) for which we want to build the network. Then, we add nodes in the "paper citations" layer of the network for the publications in the selected journal (or patents in the class), in the order of publication date, after which we draw the directed citation edges to the already added nodes[3]. In parallel to adding a knowledge artifact node, we draw inter-layer authorship edges between the artifact node and the nodes representing disambiguated authors. We add a node for the author that is not yet in the network, i.e., authors are added to the network when they first publish. The resulting network is illustrated in Fig. 2.2. The figure shows the network in two consecutive time steps corresponding to the addition of a new publication, i.e., to a *growth event* in the network. Specifically, publication k is added to the network together with the citations towards publications i and j. Authorship relations of the three authors are also added. Authors α and β were already in the network before publication k was added, as they had previously authored publication j. In contrast, publication k is the first for author γ so she is added at the same time as the publication. Authors α and β are referred to as *incumbents* and γ as a *newcomer* in the literature on the growth of collaboration networks [74, 86].

Let us now specify the notation introduced in Section 2.1 for our collaborative knowledge networks. A collaborative knowledge network shown in Fig. 2.2 is a two-layer network. One layer comprises of knowledge artifact nodes V^p and the citation edges E^{pc} among them. We call this layer $G^{pc} = (V^p, E^{pc})$. The other layer comprises author nodes V^a and no edges. The inter-layer edges between these two layers represent the authorship relations between the authors and publications. We refer to the bipartite network corresponding to these inter-layer edges as $G^a = (V^a, V^p, E^a)$. Without loss of generality, we assume that the inter-layer edges are directed, pointing from the author

[3] There are some citations in the data that go from an older publication to a newer one. However, these are extremely rare and we ignore them.

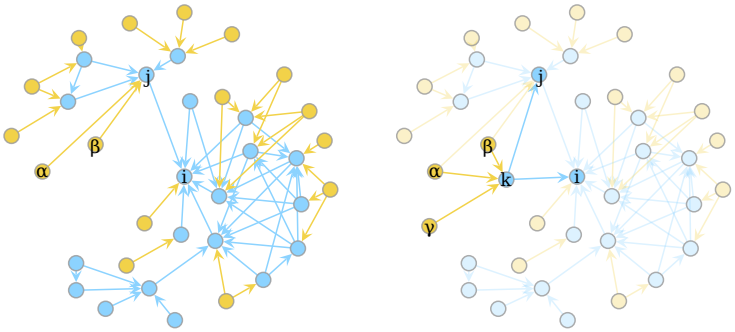


Figure 2.2: A small sample from a collaborative knowledge network corresponding to a physics journal. Only the edges directly observed in the data are shown. Blue nodes represent publications, yellow nodes represent authors. Blue edges are citations, yellow edges are authorships. (Left) the network before publication k is added, (right) publication k is added, along with its authors and citations to other publications.

to the publication. This gives the following meaning to the inter-layer in-degrees and out-degrees. The in-degree shows the number of authors for a publication node and is zero for author nodes. The out-degree shows the number of publications a given author wrote and is zero for publications. Later, when aggregating these degrees (e.g., computing the average), we only count the non-trivial nodes, i.e., only the publications for in-degree and only authors for out-degree.

Based on these basic layers we create two new layers by projection, shown in Fig. 2.3. First, we construct the co-authorship layer by projecting the inter-layer authorship edges onto authors. This layer shown on the left of Fig. 2.3 is a multi-edge network to which we refer as $G^{aa} = (V^a, E^{aa})$. Next, we project the citations between publications onto the authors. We obtain a multi-edge network with self-loops, which we call $G^{ac} = (V^a, E^{ac})$. We will use other projections, such as co-citation and bibliometric coupling, which we describe in Chapter 5 within the context.

Co-authorship as proxy of collaboration As mentioned in Chapter 1, co-authorships are used to proxy collaborations due to the lack of alternatives. However, one must exercise caution when extending the findings based on co-authorship relations to collaborations in general. There has been a growing trend towards longer author lists in the recent years [52, 128]. What does this imply about the contributions of individual authors and the collaborative efforts among them? Yitzhaki finds only a

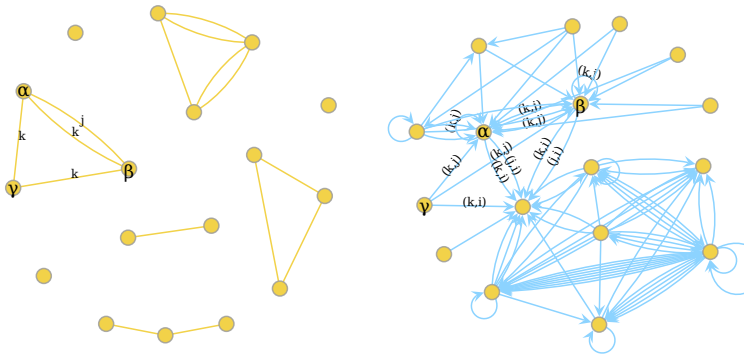


Figure 2.3: Projections of the network shown in Fig. 2.2 onto authors showing the (left) the citation relations and (right) the co-authorship relations between them.

moderately positive correlation or no correlation between the "informativeness" of a publication title and the number of authors in most of the studied cases [214]. A study by Pravdić and Oluić-Vuković investigates the relation between the productivity of authors (measured by the number of publications) and coauthorship patterns among them. They find that the mere number of coauthors is not sufficient to explain the productivity level. Instead, the "quality" of coauthorship edges must be accounted for. The frequency of coauthorship edge with the same collaborator and the type of the edge (e.g., supervisor–assistant, inter-organisational, international) affect the quality. With respect to relative contributions of different authors, Drenth speculates that more senior authors have more influence on the decisions about authorship and abuse this influence at the expense of junior researchers [52]. As an extreme of such behaviour, Cronin introduces the term "hyperauthorship" to describe the practice of having a list of more than a hundred authors [48].

Ringelmann effect states that individual productivity tends to decrease as the size of the team increases. It has been shown to affect teams in different contexts, such as collaborative software development [170]. One explanation for this decrease is the coordination cost, which increases with the team size [94, 170]. It is reasonable to assume that team of authors writing a publication is also subject to this effect. There are $n(n-1)/2$ coauthorship edges for a publication with n authors, i.e., the number of coauthorship edges grows quadratically with the length of the author list. Each edge represents a possible communication channel, thus it is reasonable to assume that coordination cost also increases super-linearly with the team size. At the limit of hyperauthorship, we can safely assume that not every pair of authors communicates

directly with each other, meaning that not all coauthorship edges derived from a long author list represent real collaborations.

There have been different proposals for how to treat coauthorships in such cases [156]. A radical approach is to attribute all the credit only to the first author [43], which may be suitable for measuring authors' productivity in some cases. However, it is not at all suitable for studying collaborations because it ignores all of them. Another approach is the *adjusted*, or *fractional count*. A weight, usually the inverse of the team size, is assigned to coauthorship edges in this case.

In our analysis, we choose a different approach, as there is no consensus on how to best treat very large author lists. We analyse communities that do not suffer from the phenomenon of hyperauthorship. As Cronin argues, life sciences and high-energy physics are the fields that suffer the most. Note that we do not put emphasis on the negative connotation of the hyperauthorship, as put by Cronin, but we rather see it as the transition to the regime when co-authorship relations cannot possibly reflect individual collaborations. In our data set on high-energy physics, INSPIRE, we find publications with more than 5000 authors. As we explain in the next section, in INSPIRE data we will only analyse the journals that do not publish considerable amount of publications with an enormous author list.

2.3 Data sets

In this section, we briefly summarise the data we use in the following chapters. The three data sets were obtained in various machine-readable text formats, such as JSON, XML and CSV. To facilitate the analyses, they were pre-processed and stored in relational (`SQLite`) and graph (`Neo4j`) databases.

INSPIRE The data set is obtained from the INSPIRE online information system for High-Energy Physics developed by a collaboration of CERN, DESY, Fermilab, IHEP, and SLAC. The system replaces the previously used Invenio and SPIRES databases. It consolidates information on High-Energy Physics publications, researchers, collaborations and research data and allows access through a web interface, API and bulk downloading of metadata. The major advantage of this data set is the high quality of author disambiguation[4]. It is facilitated by personalised features, such as author

[4] See `http://inspirehep.net/info/general/project/index`

Table 2.1: Size summary of the INPIRE data. The number of nodes and edges, as well as the median in-degrees and out-degrees of the edges of each type.

Nodes	Count
Publications	1212731
Authors	116846
Journals	2370

Edges	Count	out-deg. (median)	in-deg. (median)
Citations	6713902	8	2
Authorships	1371921	2	1

profiles and paper claiming.

Table 2.1 summarises the data set. A total of 1212731 publications in 2370 journals are authored by a total of 116846 authors. There are 6713902 citations among these publications. The median number of citations given by a publication is 8 and the median number of received citations is 2. The median number of publications written by an author is 2 (out-degree in G^a) and the median number of authors of a publication is 1 (out-degree in G^a).

As mentioned above, this data set suffers from the hyperauthorship phenomenon. This is shown in Fig. 2.4. There are 1278 publications with more than a thousand authors. While this represents only 0.1% of publications, it corresponds to more than a billion coauthorship edges. These edges do not represent direct collaborations between pairs of authors, but they can have a strong effect on the statistical outcomes of the analyses. For example, Newman points out for the related SPIRES data set that the authors in High-Energy Physics have, on average, 173 coauthors, while in other fields the number ranges from 3.87 to 18.1 [138].

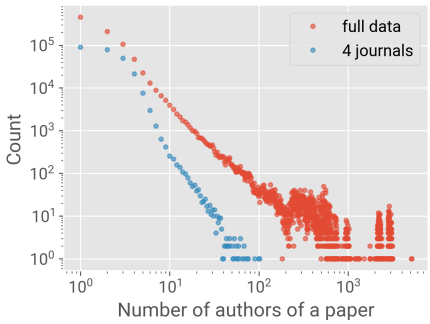


Figure 2.4: Distribution of the number of authors of a publication. Red dots correspond to the full data and the blue dots correspond to the four journals filtered by Zingg [218].

Table 2.2: Network summary for the four largest journals in the Inspire-HEP data set.

Network	$\lvert V^p \rvert$	$\lvert V^a \rvert$	$\lvert E^{pc} \rvert$	$\lvert E^a \rvert$
JHEP	15739	7994	191990	39056
PR-HEP	44829	33908	213625	115237
Phys. Lett.	22786	18078	56332	53089
Nuc. Phys.	24014	18733	125252	60018

To address the problem, Zingg proposed criteria for selecting the sub-fields in High-Energy Physics that suffer the least from hyperauthorship [218]. As the first step, the publications are identified, which are made by large and named collaborations, such as *ATLAS Collaboration* and associated with large experiments, such as *CERN-LHC-ATLAS*. Expanding this step, all publications explicitly tagged as *experimental* are identified. This is done using the additional information in the metadata, which we did not include in our data model (Fig. 2.1). These are the publications that tend to have the highest number of authors. Next, publications tagged as *theoretical* and *general physics* are identified. These generally have fewer authors. Finally, the fraction of publications corresponding to the two steps is computed in different journals, and the ones with the emphasis on the theoretical and general physics are selected. The ones with a considerable amount of publications are: The Journal of High Energy Physics (JHEP), High Energy Physics in Physical Review Journals (PR-HEP)[5], Physics Letters (Phys. Lett.), and Nuclear Physics (Nuc. Phys.). Hence, we select these four journals for our analyses. Table 2.2 summarises these by the numbers of publications $\lvert V^p \rvert$, authors $\lvert V^a \rvert$, citations among publications $\lvert E^{pc} \rvert$, and the authorship relations $\lvert E^a \rvert$.

APS journals American Physical Society has been publishing physics journals since 1893. From its foundation up to 1970, the *Physical Review* journal published articles on all of the areas of physics. The format of short letters was introduced in 1958 together with the journal *Physical Review Letters* aiming to publish notable findings from all areas of physics in a rapid fashion. With the general acceleration of publications, in 1970 *Physical Review* was split into four journals according to fields in physics. As of 2018, there are twelve journals published by APS.

APS provides the metadata on all their publications "for use in research about net-

[5] not one journal, but the collection of all publications from Physical Review family of journals that are indexed in INSPIRE data set.

Table 2.3: Size summary of the APS data. The number of nodes and edges, as well as the median in-degrees and out-degrees of the edges of each type (see Fig. 2.1).

Nodes	Count
Publications	577870
Authors	236884
Journals	13

Edges	Count	out-deg. (median)	in-deg. (median)
Citations	6713902	9	5
Authorships	1371921	2	2

works and the social aspects of science"[6]. Table 2.3 summarises the data that includes the publications up until 2015. There are 577870 publications by more than 236884 authors in 13 journals (the early Physical Review is counted separately and the two newest journals were introduced later). The reported number of authors is the lowest boundary as it corresponds to only the ones that were disambiguated by Sinatra *et al.* [181]. The upper bound is 1371921, the number of authorship relations, as prior to disambiguation each name on each publication may be considered a unique person. The disambiguation is performed by merging authors iteratively according to the following three criteria. First, the authors must have identical last names. Second, the other initials and, when provided, full first names must be the same. Third, one the following must hold: (i) there is at least one citation between the two authors, (ii) the two have at least one common (already merged) coauthor, (iii) the two authors have a similar affiliation. Sinatra *et al.* test the quality of the disambiguation based on a selection of 200 pairs of publications identified as written by the same author and 200 pairs that are identified to be written by different authors. For these 400 publications, they manually check whether the outcome of the disambiguation is correct by searching for the webpage, Google Scholar profile, if any, affiliations, etc. As a result, they find a 2% false positive rate—rate of merging two authors by mistake—and a 12% false negative rate—the rate of mistakenly splitting the author.

As with the INSPIRE data set, we construct the collaborative knowledge networks limited to specific journals. We select the following five: Physical Review (PR), Physical Review A (PRA), Physical Review C (PRC), Physical Review E (PRE), and Reviews of Modern Physics (RMP). The first covers the period from 1893 to 1970 and the others start after 1970. RMP is special in this selection as it focuses on review papers, i.e., publications that consolidate the current state of a research topic instead of contributing original research outcomes. Table 2.4 summarises these networks.

[6] `https://journals.aps.org/datasets`

Table 2.4: Network summary for five journals published by the APS.

Network	$\|V^p\|$	$\|V^a\|$	$\|E^{pc}\|$	$\|E^a\|$
PR	46728	24307	253312	87386
PRA	69147	41428	416639	144806
PRC	36039	22672	253948	108844
PRE	49118	36382	182701	95796
RMP	3006	3788	5282	5044

Patents The data set on the patents granted by the United States Patent and Trademark Office is provided by Li *et al.* [114]. The unique feature of this data set is the disambiguation of inventors, which is performed with a method called *Author-ity* [197]. The procedure is rather sophisticated, but in its core it uses statistical classification techniques for deciding whether two names on two given patents refer to the same inventor. To estimate the quality of the disambiguation, Li *et al.* compare its outcome with a curated data set of 95 prolific inventors with the total of 1169 "inventorship" relations (inventor-patent edge, similar to authorship relation). Furthermore, they conduct interviews with some of these inventors to confirm the list of their patents. The outcome of this validation is the following. The rate of false positives, i.e., wrongly matching two inventors as one, is 2.34% of inventorship relations and 2 out of 95 inventors. The false negatives, i.e, wrongly representing one inventor as two, happen in 3.26% of inventorship relations, but for 22 out of 95 inventors. Although the algorithm is far from accurate for these 95 inventors, it must be noted that these are especially prolific inventors with many patents, meaning that there is a higher than average chance to split them. Thus, this rate of false negatives estimates the upper bound of the error in the whole data set. Table 2.5 summarises the data set. A total of 4243972 patents are authored by 2703567 disambiguated inventors, with the median of 2 inventors per patent and 1 patent per inventor. There are 33533030 citations between patents, with the median of 4 outgoing and 3 incoming citations per patent. The patents are also classified according to United States Patent Classification (USPC), which is a very detailed scheme at 201231 unique classes, with the median of 27 patents per class. A patent can be attributed to multiple classes, leading to the median of 3 classes per patent in the data set.

Similar to scientific publications, we consider subsets of patents data defined thematically. We choose three USPC classes in different industries. For following classes, Table 2.6 provides the summary of the resulting networks:

Table 2.5: Size summary of the patents data. The number of nodes and edges, as well as the median in-degrees and out-degrees of the edges of each type.

Nodes	Count
Patent	4243972
Inventor	2703567
Class	201231

Edges	Count	out-deg. (median)	in-deg. (median)
Citations	33533030	4	3
Inventorship	9358182	1	2
Classification	17642776	3	27

Table 2.6: Network summary for the three patent classes.

Network	$\lvert V^p \rvert$	$\lvert V^a \rvert$	$\lvert E^{pc} \rvert$	$\lvert E^a \rvert$
PAT 320	8199	10323	45741	17215
PAT 424	8266	13010	14927	19680
PAT 703	10098	18548	22249	25533

- Patent Class 320 "Data processing: structural design, modeling, simulation, and emulation" (PAT 320)
- Patent Class 424 "Drug, bio-affecting and body treating compositions" (PAT 424)
- Patent Class 703 "Electricity: battery or capacitor charging or discharging" (PAT 703)

2.4 Methods

Behind all of the contributions in this dissertation are statistical methods of network theory. Namely, statistical network ensembles are utilised as null models in Part I and generative growth models underlie Part II. Moreover, general methods from statistics are extensively used to summarise and interpret the outcomes of network analyses and to derive stylised facts about the collaborative knowledge networks.

2.4.1 Network ensembles

In Section 1.2 we briefly introduced the notion of a network ensemble and specifically, the *configuration model*. Here we provide a formal definition of ensembles and of generative network models.

Let us start with some fixed aggregate network statistic X. We choose a generative model $\mathcal{M}(X)$ producing network realisations characterised by the aggregate statistic X. We denote the set of all network realisations G characterised by the statistic X as the sample space $\mathcal{W}(X)$. A probability $\Pr(G)$ can be assigned to each $G \in \mathcal{W}(X)$, which is usually defined by or derived from the generative model $\mathcal{M}(X)$. Then, the ensemble $\mathcal{E}(\mathcal{M}, X)$ is the set of all network realisations resulting from $\mathcal{M}$ and having the aggregate statistic X, together with the probability to observe each of these realisations.

Erdös-Rényi model The aggregate statistic X can be as simple as the number of nodes n and edges m. The ensemble defined by these two numbers is often denoted as $G(n, m)$. It is a variant of the *Erdös-Rényi model* [63], which is originally defined as a model of simple graphs with n nodes and a fixed probability p for each of the $n(n-1)/2$ possible edges to be drawn, and is denoted as $G(n, p)$. One can verify that the size of the sample space is

$$|\mathcal{W}(n, m)| = \binom{\binom{n}{2}}{m}; \quad P(G) = \frac{1}{|\mathcal{W}(n, m)|}. \tag{2.8}$$

The probability $P(G)$ to find a specific network is a constant, because every random realization of a network is equiprobable. The correspondence between the $G(n, m)$ and $G(n, p)$ models is lies in setting the *expected number of edges* $\langle m \rangle$ instead of a fixed m as

$$\langle m \rangle = p \binom{n}{2}; \quad P(G) = p^m \cdot (1-p)^{\binom{n}{2}-m}. \tag{2.9}$$

The probability $P(G)$ to find a specific network with precisely m edges is now no longer a constant, but varies with m. One can verify that $P(G)$ is maximum if $m = \langle m \rangle$. Also, the sample space $\mathcal{W}(n, p)$ of the respective networks has largely *increased* by going over from a fixed m to a fixed expected value $\langle m \rangle$.

Configuration model Which other constrains can we use to define the ensemble of networks? In the configuration model [129] introduced by Molloy and Reed, in addition to n and m, the degree sequence $K(G)$ of nodes are also fixed. The degrees of the nodes resulting from such sequence are illustrated by "stubs" for half-edges, as shown in Fig. 1.2, because a degree is defined by one of the two endpoints of edges.

In the configuration model, these stubs are preserved. However, the second node to which an edge connects can be randomly chosen with respect to the available stubs of other nodes. This leads to a different sample space $\mathcal{W}(K)$ defined by the given degree sequence K which is also much smaller than it is for the $G(n,m)$ model. But just as with $G(n,m)$, the probability for a network realisation by the configuration model is the same for all realizations, $P(G) = 1/|\mathcal{W}(K)|$. The configuration model can be defined for both simple and multi-edge networks, as well as for both directed and undirected networks. Similar to Erdös-Rényi model, the original Molloy-Reed configuration model can be extended to assume only the *expected degree sequence* $\langle K \rangle$, which we discuss in Chapter 3. Note, that for both the configuration model and the Erdös-Rényi model, the probability of a network realisation is equivalent to the probability of the corresponding adjacency matrix,

$$\Pr(G) = \Pr(\mathbf{A}). \tag{2.10}$$

This equivalence holds for any model of static networks.

Growth models A class of generative network models is designed for studying growing networks. In many of these models, new nodes and edges are sequentially added to the network according to certain mechanisms. Often, nodes are added to the network one at a time, together with a number of edges originating at the new node. The simplest probabilistic mechanism would be to connect these edges uniformly at random to the nodes already in the network. However, many real-world networks exhibit evidence of *proportional growth* [1, 157, 180, 193]. In such growth, the probability of a new edge to connect to a given node scales with the number of edges the node already has. Such models for networks are commonly called *preferential attachment models*. The case when the probability is linear on the current degree of a node is called Barabási-Albert model [9]. So, for a new node i added at the time $t+1$, the probability to connect to node j with degree $k_j(t)$ is

$$P_j(t+1) = \Pr\big((i,j); G(t)\big) = \frac{k_j(t)}{\sum_{l \in V(t)} k_l(t)}. \tag{2.11}$$

The model is defined for directed networks as well, where the preference is defined according to the in-degree of a node

$$P_j(t+1) = \frac{k_j^{\text{in}}}{\sum_{l \in V(t)} k_l^{\text{in}}(t)}. \tag{2.12}$$

There are multiple variants of preferential attachment models that make it suitable for a range of phenomena in network growth. The *fitness model* [14] introduces heterogeneity in the nodes in terms of an intrinsic ability to compete for new edges. This is described by a constant parameter η_i for each node i. This parameter then scales the probability for new edges to connect to a given node as

$$P_j(t+1) = \frac{\eta_i k_j(t)}{\sum_{l \in V(t)} \eta_l k_l(t)}. \tag{2.13}$$

Another variant of preferential attachment introduces *ageing*, or *relevance decay*, of nodes [125]. The added assumption is that older nodes lose their attractiveness for new edges over time. This relevance of the node j is introduces as a time-dependent *relevance parameter*, $R_j(t)$. The probability of the newly drawn edge from i to connect to a j is given by

$$P_j(t+1) = \frac{k_j(t)R_j(t)}{\sum_{l \in V(t)} k_l(t)R_l(t)}. \tag{2.14}$$

We will apply this model, among others, to study the growth of collaborative knowledge networks in Part II because it is a reasonable—if not a necessary—assumption that knowledge artifacts lose their relevance over time (studied in detail in Chapter 7).

2.4.2 Statistics

Most of the generative models we use to study the networks are *parametric*, meaning that they are described by one or more parameters, such as the η_i parameter in Eq. (2.13) of the fitness model. These parameters are not known and their best values must be *inferred* by comparing the model and the observed data. Moreover, often there are multiple competing models for explaining the observed data. To select the best one, there exist various *model selection* techniques.

Maximum Likelihood Estimation (MLE) Many inference methods in statistics are based on MLE. The principle is to find the parameters of the given model that maximize the likelihood to observe the available data according to the model. A parametric statistical model $\mathcal{M}(\theta)$ corresponds to a certain probability distribution $f_{\mathcal{M}}(x;\theta)$, where x is a random variable and $\theta \in \Theta$ is the (vector) parameter of the model. If the data are independent and identically distributed, the *likelihood function* written in logarithmic form is

$$\log \mathcal{L}_{\mathcal{M}}(\theta; x) = \sum_i f_{\mathcal{M}}(x_i;\theta). \tag{2.15}$$

Then, the parameter maximizing the likelihood is taken,

$$\hat{\theta} = \arg\max_{\theta} \mathcal{L}_{\mathcal{M}}(\theta|\mathbf{x}) \tag{2.16}$$

Maximum likelihood estimation is one of the simplest yet very powerful tools in statistics for fitting models to data. From Bayesian statistics viewpoint, MLE is too restrictive because it assumes uniform distribution of the parameters and is a special case of a Bayesian estimator. Also, it is a point estimator, meaning that it gives the single value of the parameters that maximizes the likelihood but does not give the confidence intervals. However, if the distribution $f(x;\theta)$ satisfies certain conditions, the estimate of the parameters is normally distributed with a known covariance matrix, which means confidence intervals of the parameters can be calculated.

Linear regression In order to model the relation between a *dependent variable* y and *explanatory variables* $\mathbf{x} = \{x_1, \ldots, x_q\}$, in discriminative statistics the *linear regression* is usually the first method to try. As the name suggests, the method models the linear relationship between the variables,

$$y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_q x_{q,i} + \varepsilon_i, \qquad i \in 1, \ldots, N \tag{2.17}$$

where the subscript i denotes a single observation in the data of size N, $\beta_0, \beta_1, \ldots, \beta_q$ are the parameters of the model and ε is the *error term*, which accounts for the stochastic noise. Linear regression is also used to model non-linear relations between variables by transforming the variables first. For example, exponential relation is modelled by a linear regression between log-transformed variables.

There are different methods for fitting the linear model. A standard one is the *ordinary*

Table 2.7: Significance codes of parameter estimates used throughout the dissertation.

p-value	< 0.001	< 0.01	< 0.05	< 0.1
Signif. code	***	**	*	.

least squares, which minimises the sum of square *residuals*—the difference between the observed values of the dependent variable and the values predicted by the model. For this methods to produce reliable outcomes, to be *valid*, certain conditions must be met. First, the error term must be normally distributed, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Second, the variance of the error must not depend on the explanatory variables $\mathrm{Var}(\varepsilon \mid \mathbf{x}) = \sigma^2$, a condition known as "homoskedasticity". Third, the expectation of the error term given the explanatory variables must be zero, $\mathrm{E}(\varepsilon \mid \mathbf{x}) = 0$.

The predictive quality of the linear regression model is commonly assessed be means of the *coefficient of determination* R^2. It denotes the fraction of the variance of the dependent variable that is predictable by the exploratory variables, given the regression model. The coefficient of determination is defined as

$$R^2 = \frac{\sum_1^N (\bar{y}_i - \langle y_i \rangle)^2}{\sum_1^N (y_i - \langle y_i \rangle)^2}, \tag{2.18}$$

where $\bar{y}_i$ are the predicted values of the dependent variable by the regression model, y_i are the observed values and $\langle y_i \rangle$ is their mean.

Linear regression provides p-values for the estimates of the parameters β_l. In general, for a statistical model the p-value is the probability that the magnitude of a certain *statistic* under the *null hypothesis* is greater or equal than under the *alternative hypothesis*. In linear regression, the p-value of each parameter is calculated according to the null hypothesis that the parameter is zero, i.e., the corresponding explanatory variable has no effect, and the alternative that the parameter has the value estimated by the regression analysis. The null hypothesis is rejected for low p-values. The opposite is not true, the null hypothesis is not accepted if the p-value is high. So, if the regression analysis results in low p-value for a given parameter, then we say that the parameter and the effect of the corresponding explanatory variable is *significant*. The interpretation of the p-value is similar for other statistical methods of model selection (see below) and parameter estimation. There are widely accepted heuristics for the *significance level*, that sets the threshold for the p-value, below which the null hypothesis is rejected. Throughout the dissertation, we will use the codes shown in

Table 2.7 next to the parameter estimates and statistical test results.

Model selection MLE is used to find the best parameters for a given model to explain the data. However, on its own it is not suitable when different models need to be compared. In this case, a common approach is to compare the models in terms of *information criteria*. The most common ones—*Akaike information criterion (AIC)* [4] and *Bayesian information criterion (BIC)* [173]—are based on MLE and, in addition, penalize for *model complexity* (the degrees of freedom of the model, which is often equal to the number of parameters). That means, if two competing models lead to the same maximum likelihood value, then the one that has lower model complexity, i.e., less degrees of freedom, is preferred. For the model $\mathcal{M}$ with the number of parameters $|\mathcal{M}|$,

$$AIC(\mathcal{M}) = -2\log\mathcal{L}(\hat{\theta}) + 2|\mathcal{M}|, \tag{2.19}$$

$$BIC(\mathcal{M}) = -2\log\mathcal{L}(\hat{\theta}) + 2|\mathcal{M}|\log N. \tag{2.20}$$

According to both criteria, the model with a higher likelihood and a smaller number of parameters is preferred, meaning that the smallest AIC is the criterion for model selection. There are certain conditions that the models must fulfil for the two criteria to be reliable. Generally, AIC is less restrictive on the models. In particular, BIC applies only to models that are *nested*, i.e., one model can be reduced to another by removing some of the parameters, while AIC does not require such nestedness. For this reason, we will use AIC in Chapter 6.

Instead of only taking the model with the lowest AIC, the model selection can be done in a more principled way that gives weights to different models. For example, consider that in the set of models we have AIC differences in the range of thousands, but there are two models for which the AIC are almost equal. In this case, it would be wrong to select only one model. Instead, we choose to use a statistical heuristic called *relative likelihood of models*,

$$w(M) = \exp\left((\min_{\mathcal{M}}(AIC) - AIC(M))/2\right). \tag{2.21}$$

This measure is similar to the *likelihood ratio* $\Lambda(\mathbf{x})$, but for $w(M)$ to be applicable, the models that are compared do not have to be nested. *Likelihood ratio test* is a

theoretically principled test for comparing two nested models $\mathcal{M}_0$ and $\mathcal{M}_1$.

$$\Lambda(\mathcal{M}_0, \mathcal{M}_1; \mathbf{x}) = \frac{\mathcal{L}_{\mathcal{M}_0}(\theta_0 \mid \mathbf{x})}{\mathcal{L}_{\mathcal{M}_1}(\theta_1 \mid \mathbf{x})}. \tag{2.22}$$

Wilks theorem states that the test statistic $-2\log(\Lambda)$ follows the $\chi^2(|\mathcal{M}_1| - |\mathcal{M}_0|)$ [209], meaning that the significance of the advantage of one model against the other can be computed. We will apply the likelihood ratio test in Chapter 3 for the testing of hypotheses about the structural properties of networks.

Part I

Structure

We cannot tell the precise moment when friendship is formed. As in filling a vessel drop by drop, there is at last a drop which makes it run over, so in a series of kindnesses there is at last one which makes the heart run over.

James Boswell

Chapter 3

Generalized hypergeometric network ensembles

Summary

Statistical ensembles of networks, i.e., probability spaces of all networks that are consistent with given aggregate statistics, have become instrumental in the analysis of complex networks. We introduce *generalized hypergeometric ensembles*, a new class of analytically tractable statistical ensembles of finite, directed or undirected, and multi-edge networks. Utilising the analytical tractability of the ensembles, we provide methods for model selection and hypothesis testing for various topological patterns. The ensembles generalise the configuration model, which is used to model networks with a given degree sequence or distribution. The generalisation rests on the introduction of dyadic *edge propensities*, which capture the degree-corrected tendencies of pairs of nodes to form edges between each other. Furthermore, we demonstrate how the ensemble can be used for the extraction of network backbone both on synthetic and empirical data.

Based on Casiraghi G., Nanumyan V., Scholtes I., and Schweitzer F., "From Relational Data to Networks: Testing Hypotheses about the Origin of Repeated Interactions", *pending submission*; Casiraghi G., Nanumyan V., Scholtes I., and Schweitzer F. (2017) "From Relational Data to Graphs: Inferring Significant Links Using Generalized Hypergeometric Ensembles", in *Social Informatics. SocInfo 2017*, Springer; Samarin M. (2016) "Modelling co-locations in human mobility", *Master's thesis*. VN conceptualized the network ensemble and had a major contribution to its formulation and applications. VN was the lead contributor to the network reconstruction. While supervising the Master's thesis of M. Samarin, VN provided the code for building the ensemble and classifying edges.

The need for novel statistical ensembles to model social interactions behind collaborative knowledge networks was motivated in Chapter 1. We extend this motivation, as the analysis of relational data from the perspective of networks has become a key method not only in scientometrics and bibliometrics but in the study of complex systems across different domains. Examples for *network science* techniques include (i) algorithms to detect cluster or community structures, (ii) quantitative measures capturing the importance, or centrality, of nodes, (iii) statistical methods to detect significant patterns such as frequent sub-graphs or motifs, and (iv) techniques to study the evolution (or control) of dynamical processes in complex networks. Their application to relational data that capture interactions between elements of complex social, biological or technical systems has become popular. However, despite this popularity, we still lack methods to answer a crucial question: *when is it justified to use graph or network models to study relational data?* Depending on the characteristics of real-world data sets, this question has multiple facets that pose severe challenges to contemporary network science.

First, relational data on networked systems are increasingly *time-resolved* or *sequential*, i.e., we know *when* or *in which order* relations occurred. The application of standard (static) network modelling techniques to such data discards information on temporal correlations and can yield wrong results, e.g., about community structures, node centralities or dynamical processes [120, 153, 169, 172]. Secondly, different from examples like telecommunication, transportation, or citation networks where the interpretation of an *edge* is straightforward, inferring the *network topology* of a system is often a non-trivial problem by itself. This is because relational data often capture observations of *repeated interactions* between elements, which may or may not be direct expressions of an explicit underlying network topology. For instance, recall the problem of representing collaborations with coauthorship edges discussed in Section 2.2—one of many examples of modelling multiple observations of *interactions* between agents in a social system as a *social network* [10, 55, 56]. Similarly, data on the *co-expression* of genes is used to construct *biological networks*, or the *co-occurrence* of words or concepts is used to construct semantic networks in language studies. The question whether or not such observed interactions justify the hypothesis of a non-trivial underlying network topology must be answered *before* applying any network-analytic methods.

To illustrate this problem, consider the toy example shown in Fig. 3.1 (a). Here, we observe a total of 35 repeated interactions between four nodes A, B, C and D, captured as a set of triples (i, j, n_{ij}) where n_{ij} counts the number of times i and j interacted. Let us

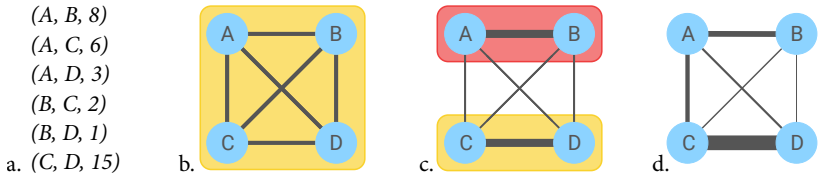


Figure 3.1: Illustrative example of data on repeated interactions and three different types of (network) models for these interactions: the first model (a) accounts for the rates at which nodes engage in interactions while disregarding the topology of these interactions. The second model (b) additionally accounts for a group structure that influences interaction rates. The third model (c) fits both the weight and the direction of edges to the observed interactions. Answering which of these models should be used to study a given data set on repeated interactions is a model selection problem that must precede the analysis of relational data from a network perspective.

further consider three different hypotheses about the mechanisms driving these interactions. First, we could explain the frequency of interactions between pairs of nodes simply by the configuration model. To recall, it only respects *how often* each node has been the source or the target of an interaction with *any* other node. This simple *mean field model*, illustrated in Fig. 3.1 (b), does not assume an underlying network topology, i.e., nodes have no intrinsic preferences with *whom* they interact. Considering that all nodes can potentially interact with each other, it rather assumes that both the frequency and topology of observed interactions result from the activities of nodes. We can augment this model by additional hypotheses such as, e.g., the presence of group structures that influence the interactions between nodes. An illustration of such a *group model* is shown in Fig. 3.1 (c). Here, we assume that nodes have a preference to interact with other nodes in the same group, but not with any specific nodes. Finally, we can consider a *network model* for the observed interactions, where weighted edges of the network determine both the topology and the frequency of interactions between *particular pairs* of nodes. Such a model, illustrated in Fig. 3.1 (d), corresponds to a straightforward interpretation of observed interactions as a *weighted network*. That is, we assume that the topology and frequency of interactions are direct expressions of a network structure that captures the strength of interaction preferences between nodes. This network model clearly offers the best explanation for the observed interactions. However, it is also *a maximally complex model*, as it simply encodes the phenomenon that it seeks to explain.

This simple example illustrates an important problem. In settings where observed interactions do not directly correspond to an underlying network topology, deciding *how* these interactions should be modelled is not trivial. This questions the naive

modelling of data from the perspective of graphs or complex networks and highlights that the construction of graph or network abstractions based on observational data is a difficult inference task. The importance of this problem has recently been acknowledged in data mining, bioinformatics, and network science, and it has been addressed from the perspective of *link prediction* [115], *network inference* [5], or *graph identification* [132]. However, most of these works start from the assumption that there *is* an underlying graph that drives observed interactions. We still lack principled approaches to (i) test whether a network model is justified, and (ii) contrast this hypothesis with alternative explanations for the data. Closing this research gap requires model selection techniques that take into account both the *complexity of a model* and its *explanatory power* for the observed interactions. The availability of such techniques is key to answering the crucial question of *when* a network abstraction of relational data is justified, and when we can reasonably apply network science techniques [28, 169, 182].

To address this issue, we introduce a statistical modelling framework that (i) allows formulating a wide range of different hypotheses about the origin of repeated interactions, (ii) provides statistically principled techniques to test these hypotheses in empirical data, and (iii) allows inferring the significant relations in data on noisy interactions.

3.1 Ensemble formulation

Let us consider empirical data consisting of repeated interactions (i, j) between nodes i and j. As described in Section 2.1, such relational data can be represented as a *multi-edge* network $\hat{G} = (V, E)$, with a set V of n nodes, and a multi-set $E \subseteq V \times V$ of m (directed) edges. In the following, we use a "hat" notation to characterise the empirical network. For instance, its integer-valued adjacency matrix writes as $\hat{\mathbf{A}}$.

Configuration model Our construction of a statistical ensemble follows the idea of the Molloy-Reed configuration model [129], which is to randomly shuffle the topology of a network G while preserving node degrees, as already described in Section 2.4.1. The configuration model uses a *node-centric sampling* approach, generating edges between randomly sampled pairs of nodes such that the *exact* observed degrees of nodes are preserved. Different from this, we utilize an *edge-centric sampling* of m edges from the set of all possible edges such that the sequence of *expected* degrees of nodes is preserved.

We first define a matrix $\boldsymbol{\Xi}$, which sets the maximum possible number Ξ_{ij} of multi-edges that can exist between each pair of nodes i and j as $\Xi_{ij} = \hat{k}_i^{\text{out}}\hat{k}_j^{\text{in}}$ [100, 140]. We can hence define a statistical ensemble based on the following generative model. We sample edges from the n^2 sets Ξ_{ij} of possible multi-edges uniformly at random. This can be viewed as an *urn problem* [96] where edges to be sampled are represented by balls in an urn. We specifically obtain an urn with $M = \sum_{i,j}\Xi_{ij}$ balls having $n^2 = |V \times V|$ different colours, each colour representing all possible edges between a given pair of nodes. The sampling of a network corresponds then to drawing exactly m balls from this urn. Each adjacency matrix $\mathbf{A}$ with $\sum_{i,j}A_{ij} = m$ corresponds to one particular realisation drawn from this ensemble. The probability to draw exactly A_{ij} edges between all pair of nodes $i, j \in V$ is given by the *multivariate*[1] hypergeometric distribution. To be precise, the multivariate variable described by the distribution is obtained by stacking the rows of the $n \times n$ adjacency matrix $\mathbf{A}$ into a $n^2 \times 1$ vector. For simplicity of notation and without loss of generality, we do not distinguish between the matrix and its stacked vector representation. Hence,

$$\Pr(\mathbf{A}) = \binom{M}{m}^{-1}\prod_{i,j}\binom{\Xi_{ij}}{A_{ij}}, \tag{3.1}$$

which provides an analytical expression for the probability of the given corresponding network G.

For each pair of nodes $i, j \in V$, the probability to draw exactly $\hat{A}_{ij}$ edges is given by the marginal distribution of the multivariate hypergeometric distribution

$$\Pr(A_{ij} = \hat{A}_{ij}) = \binom{M}{m}^{-1}\binom{\Xi_{ij}}{\hat{A}_{ij}}\binom{M-\Xi_{ij}}{m-\hat{A}_{ij}}. \tag{3.2}$$

We can further calculate the expected number of edges between any pair of nodes i and j as $\langle A_{ij}\rangle = m\frac{\Xi_{ij}}{M}$. This leads to the expected in-degrees (out-degrees) of all nodes

[1] of multiple variables

by summing the rows (columns) of matrix $\langle A_{ij} \rangle$:

$$\langle k_j^{\text{in}} \rangle = \sum_{i \in V} \langle A_{ij} \rangle = m \frac{\sum_{i \in V} \hat{k}_i^{\text{out}} \hat{k}_j^{\text{in}}}{M} = \hat{k}_j^{\text{in}}, \tag{3.3}$$

$$\langle k_j^{\text{out}} \rangle = \sum_{i \in V} \langle A_{ji} \rangle = m \frac{\sum_{i \in V} \hat{k}_j^{\text{out}} \hat{k}_i^{\text{in}}}{M} = \hat{k}_j^{\text{out}}. \tag{3.4}$$

Equation (3.4) confirms that the *expected* in-degree and out-degree sequences of realisations drawn from the resulting statistical ensemble correspond to the degree sequences of the given network $\hat{G}$. We thus arrive at a *hypergeometric network ensemble*, which (i) provides a generalization of the configuration model for directed, multi-edge networks, (ii) has a fixed sequence of *expected* degrees, and (iii) is analytically tractable. The above formulation for directed networks can be adapted to undirected networks, to networks with and without self-loops. Furthermore, we obtain a framework for the generalization of other generative models like, e.g., the multi-edge version of the Erdös-Rényi model [63], where only n and m are fixed, while there are no constraints on the degree sequence. That is achieved by setting $\Xi_{ij} = m^2/n^2 = \text{const}$, which directly results from $\langle k_i^{\text{in}} \rangle = \langle k_i^{\text{out}} \rangle = m/n$.

The sampling procedure outlined above provides a parsimonious stochastic model for multi-edge, directed networks in which (i) the expected in-degrees and out-degrees of nodes are fixed, and (ii) edges between these nodes are generated at random. This stochastic model can serve as a *null model* that considers *combinatorial effects* and *no* additional correlations. That means, if a given network does not significantly differ from this null model, we learn that the interactions behind that network are only driven by the *activities* (degrees) of the nodes. If, however, the network cannot be described by the model, the question is to what extent the patterns in a given empirical network exhibit statistically significant *deviations* from this null model (addressed in Section 3.2).

Edge propensity To answer this question, we generalise the *hypergeometric ensemble* as follows. We introduce a matrix $\mathbf{\Omega}$ whose entries Ω_{ij} capture relative *edge propensities*, i.e., the tendency of a node i to form an edge *specifically* to node j. In particular, we assume that an entry Ω_{ij} captures the propensity that goes *beyond* the tendency of a node i to connect to a node j that results from combinatorial effects,

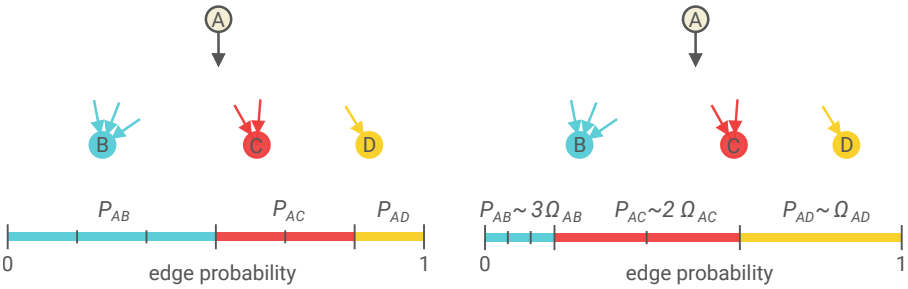


Figure 3.2: The probabilities for the node A to connect to nodes B, C and D according to (left) the configuration model (or unbiased hypergeometric ensemble) and (right) the hypergeometric ensemble with different propensities $\Omega_{AB} < \Omega_{AC} < \Omega_{AD}$. Even though in configuration model the edge (A, B) has three times the probability of (A, D), the latter is more likely when the propensity is accounted for.

i.e., a *degree-corrected preference* of i linking to j which accounts for the in-degree of j and the out-degree of i. The key idea of our generalised ensemble is to use the edge propensities Ω_{ij} to *bias* the edge sampling process described above. Similar to the urn model, biased sampling implies that the probability of drawing balls of a given colour does not only depend on their number but also on the respective relative propensities. This is illustrated in the left panel of Fig. 3.2, where the probabilities of the outgoing stub of node A to connect to nodes B, C, D are shown according to configuration model. The probability of each node to be selected is proportional to the respective number of available stubs. The effect of edge propensities Ω_{Aj} is to scale the probability of each corresponding stub of j to be selected, as shown on the right of Fig. 3.2. The probability distribution resulting from such a biased sampling process is given by the multivariate *Wallenius' non-central hypergeometric distribution* [40, 68, 203]:

$$\Pr(\mathbf{A}) = \left[\prod_{i,j}\binom{\Xi_{ij}}{A_{ij}}\right]\int_0^1 \prod_{i,j}\left(1 - z^{\frac{\Omega_{ij}}{S_\Omega}}\right)^{A_{ij}} dz \tag{3.5}$$

with $S_\Omega = \sum_{i,j}\Omega_{ij}(\Xi_{ij} - A_{ij})$.

Similar to unbiased sampling, the probability to observe a particular number $\hat{A}_{ij}$ of edges between a pair of nodes i and j can again be calculated from the marginal

distribution:

$$\Pr(A_{ij} = \hat{A}_{ij}) = \binom{\Xi_{ij}}{\hat{A}_{ij}} \binom{M - \Xi_{ij}}{m - \hat{A}_{ij}} \cdot \int_0^1 \left[\left(1 - z^{\frac{\Omega_{ij}}{S_\Omega}} \right)^{\hat{A}_{ij}} \left(1 - z^{\frac{\bar{\Omega}_{\setminus(i,j)}}{S_\Omega}} \right)^{m - \hat{A}_{ij}} \right] dz \tag{3.6}$$

where $\bar{\Omega}_{\setminus(i,j)} = (M - \Xi_{ij})^{-1} \sum_{(l,m)\neq(i,j)} \Xi_{lm}\Omega_{lm}$.

The entries of the expected adjacency matrix $\langle A_{ij} \rangle$ can be obtained by solving the system of equations described in [68]. Note that for the special case of a uniform edge propensity matrix $\mathbf{\Omega} \equiv$ const, which corresponds to an unbiased sampling of edges, the integral in Eq. (3.5) becomes $\binom{M}{m}^{-1}$ and we thus recover Eq. (3.1) [203].

A major advantage of the formalism outlined above is that, by specifying different edge propensities matrices $\mathbf{\Omega}$, we obtain a broad class of *generalised hypergeometric network ensembles*. This allows us to encode a wide range of dyadic patterns in networks, while still obtaining an analytically tractable statistical ensemble corresponding to a simple and well-defined generative model. Moreover, the ensemble allows fit the propensities to a given network. For this, we use the property of the Wallenius' distribution for the expected adjacency matrix $\langle A_{ij} \rangle$, according to which

$$\left(1 - \frac{\langle A_{11} \rangle}{\Xi_{11}} \right)^{\frac{1}{\Omega_{11}}} = \left(1 - \frac{\langle A_{12} \rangle}{\Xi_{12}} \right)^{\frac{1}{\Omega_{12}}} = \dots \tag{3.7}$$

with the constraint $\sum_{i,j} \langle A_{ij} \rangle = m$. Assuming that the observed edges *are* the expected ones, we can solve the above equation for Ω_{ij} to obtain the best fitting propensities. So, solving the linear system

$$\begin{cases} \left(1 - \frac{\langle A_{11} \rangle}{\Xi_{11}} \right)^{\frac{1}{\Omega_{11}}} & = C, \\ \left(1 - \frac{\langle A_{12} \rangle}{\Xi_{12}} \right)^{\frac{1}{\Omega_{12}}} & = C, \\ & \vdots, \end{cases} \tag{3.8}$$

we obtain the fitted propensities up to an arbitrary negative constant $k \in \mathbb{R}^-$ as follows

$$\Omega_{ij} = \frac{1}{k} \log \left(1 - \hat{A}_{ij}/\Xi_{ij} \right) \quad \forall i, j \in V. \tag{3.9}$$

In the following, we show how the class of generalised hypergeometric ensembles can be used for *model selection* and *hypothesis testing* in complex networks.

3.2 Model selection and hypothesis testing

In this section, we present two statistical tools for model selection and hypothesis testing in networks: (i) the *likelihood-ratio test* allows to compare nested hypotheses and (ii) *Mahalanobis distance* shows how far an observation is from a distribution, effectively providing a goodness-of-fit for the model.

Likelihood-ratio test Let H_r be some statistical hypothesis about interactions. Here we always assume that each hypothesis is defined by a hypergeometric model and can be encoded by some propensity matrix $\mathbf{\Omega}_r$. Each $\mathbf{\Omega}_r$ is characterized by a number of free parameters that we want to fit to the data, such that the probability to observe the data is maximized. Given a propensity matrix $\mathbf{\Omega}_r$, we can compute the probability of observing the data $\mathbf{A}$ as $\Pr(\mathbf{A} \mid \mathbf{\Omega}_r)$, where the probability is computed as in Eq. (3.5). The likelihood $L_r(\mathbf{\Omega}_r|\mathbf{A})$ of the propensity matrix $\mathbf{\Omega}_r$ is then equal to $\Pr(\mathbf{A} \mid \mathbf{\Omega}_r)$ in the frequentist approach, which in Bayesian approach assumes a flat prior distribution of $\mathbf{\Omega}_r$. We indicate the propensity matrix corresponding to the maximum likelihood estimation of the free parameters in $\mathbf{\Omega}_r$ as $\bar{\mathbf{\Omega}}_r$. We refer to [35] for the actual estimation procedure.

The number of free parameters in each propensity matrix defines the number of degrees of freedom of each statistical hypothesis. The number of degrees of freedom of the maximally simple model $\mathbf{\Omega}_0$ corresponding to the configuration model is zero, since there is no parameter fitted to the data. The maximally complex network hypothesis ($\mathbf{\Omega}_n$) obtained using Eq. (3.9) has as many degrees of freedom as the number of allowed pairs of nodes minus one. This is because the network hypothesis corresponds to fitting a parameter for every pair of nodes such that the expectation of the model defined by $\mathbf{\Omega}_n$ coincides with the observed data. Hence, in the case of a directed network with self-loops, $\mathbf{\Omega}_n$ has $n^2 - 1$ degrees of freedom, where n is the number of nodes.

The degrees of freedom of intermediate models depend on the number of parameters fitted. In the case of a group model $\mathbf{\Omega}_g$ there is only one parameter. It defines the relative propensity for within-group interactions against across-group interactions, thus the number of degrees of freedom is one, independently of the number of groups.

If we denote the group of node i as $g(i)$,

$$\Omega_{g,ij} = \begin{cases} 1 & \text{for } g(i) = g(j), \\ c & \text{otherwise, with } c \in [0, 1]. \end{cases} \tag{3.10}$$

Assume now we have two hypotheses to test against each other. Let H_0 denote the more simple hypothesis as the null-hypothesis and let H_a denote the alternative. The corresponding propensity matrices are $\mathbf{\Omega}_0$ and $\mathbf{\Omega}_a$. To test the alternative hypothesis against the null, we use the likelihood-ratio statistic $\Lambda(0, a)$, defined as follows:

$$\Lambda(0, a) = \frac{\mathcal{L}_0(\bar{\mathbf{\Omega}}_0 \mid \mathbf{A})}{\sup(\mathcal{L}_0(\bar{\mathbf{\Omega}}_0 \mid \mathbf{A}), \mathcal{L}_a(\bar{\mathbf{\Omega}}_a|\mathbf{A}))} \tag{3.11}$$

Thanks to Wilks' theorem [209], if $\mathbf{\Omega}_0$ is a special case of $\mathbf{\Omega}_a$, i.e., the null-hypothesis corresponds to the model with fewer parameters and can be formulated by constraining the model with more parameters (the two models are nested), the distribution of $-2\log(\Lambda(0, a))$ can be approximated by the χ^2 distribution with as many degrees of freedom as the difference of degrees of freedoms between the two models [35]. If ν is the difference of degrees of freedom between the null and the alternative models, the p-value of the likelihood-ratio test between the two hypotheses can be computed as follows:

$$p = \Pr\left[\chi^2(\nu) \geq -2\log(\Lambda(0, a))\right]. \tag{3.12}$$

We then reject the null-hypothesis in favour of the alternative if the p-value is smaller than a threshold α.

The three hypotheses discussed above—the mean-field hypothesis according to configuration model, the group structure and the network hypothesis—are all nested. In fact, the mean field model can be seen as a special case of any model where all parameters are constrained to be equal. Moreover, the group model is a special case of the network model, when parameters in $\mathbf{\Omega}_n$ are constrained either to between-group or within-group parameters according to whether the corresponding pair of nodes are in the same group or not.

Mahalanobis distance and goodness of fit For the multivariate distribution in Eq. (3.5), we can use the *Mahalanobis distance* [117] to quantify how far the observed network $\hat{G}$ is from the model defined by a given $\mathbf{\Omega}_\mathbf{r}$. Better models correspond to

ensembles statistically closer to the observed data and Mahalanobis distance is the standard choice to quantify the *statistical distance* between a multivariate observation and a distribution. The (squared) Mahalanobis distance $D_r^2(\hat{\mathbf{A}})$ gives hence an estimate of how well the model defined by $\mathbf{\Omega}_r$ fits the observed data. It is the multivariate generalization of the Z-score and captures how many standard deviations an observation is away—in the corresponding direction—from the expectation. If $\langle \mathbf{A} \rangle_r$ denotes the expected adjacency matrix of the ensemble corresponding to the square of the Mahalanobis distance writes

$$D_r^2(\hat{\mathbf{A}}) = \left(\hat{\mathbf{A}} - \langle \mathbf{A} \rangle_r\right)^T \mathbf{\Sigma}_r^{-1} \left(\hat{\mathbf{A}} - \langle \mathbf{A} \rangle_r\right). \tag{3.13}$$

In Eq. (3.13) we use the vector representation of $\hat{\mathbf{A}}$ and $\langle \mathbf{A} \rangle_r$ of length n^2, with the corresponding covariance matrix $\mathbf{\Sigma}_r$ of size $n^2 \times n^2$ for the multivariate Wallenius distribution in Eq. (3.5).

For the distribution given by Eq. (3.5), $\langle \mathbf{A} \rangle_r$ can be calculated analytically and its covariance matrix $\mathbf{\Sigma}_r$ can be approximated numerically [69]. From a set of candidate models, the one with the smallest Mahalanobis distance is the best choice, since the corresponding ensemble is statistically the closest to an observed empirical network. However, such comparison of the models does not penalise directly for the model complexity.

It is also possible to assess the goodness-of-fit of a given model $\mathbf{\Omega}_r$ in absolute terms, instead of comparing to other models. A p-value of the goodness-of-fit can be given in terms of the complementary cumulative distribution $\Pr\left[D_r^2(\mathbf{A}) \geq D_r^2(\hat{\mathbf{A}})\right]$ where $\mathbf{A}$ is a random realisation drawn of the generalised hypergeometric ensemble defined by $\mathbf{\Omega}_r$. Under certain conditions, there are closed-form expressions for $\Pr\left[D_0^2(\mathbf{A}) \geq x\right]$ [2]. However, in the following we resort to a sampling procedure, which is facilitated by the simplicity of the underlying generative model.

Detecting network topology In the following we illustrate and validate our methodology using synthetically generated data on repeated interactions. These data contain a variable number of repeated interactions generated by a known (ground truth) mechanism. We specifically analyse (i) whether likelihood-ratio based model selection reliably identifies the correct model, and (ii) how many observations are needed for

[2] $\Pr\left[D_0^2(\mathbf{A}) \geq x\right]$ converges to a χ^2 distribution with $n^2 - 1$ degrees of freedom if $\Xi_{ij} \gg m$ and $\Omega_{ij} = c \; \forall i, j \in V$

it. We generate the interaction data among n nodes as follows. First, we randomly select $l < n(n-1)$ pairs of nodes i, j, $i \neq j$, and later allow interaction only between these l pairs. This effectively defines a topology underlying the interactions. Then, we randomly generate synthetic interaction data by repeatedly drawing m edges uniformly at random, with replacement, among the l pairs of nodes. Figure 3.3 shows an example of $n = 10$ nodes with $l = 11$ pairs that can interact in the top left panel, with one possible realisation of $m = 60$ edges in the right panel. The bottom panel of Fig. 3.3 shows the outcome of the likelihood-ratio test between the observed network topology and the configuration model for different values $l \in (1, n(n-1))$ and $m \in (34, 4126)$. For each combination of l and m, the test is performed 30 times—each time regenerating both the topology and the realisation of interactions—and the fraction of outcomes in favour of the network model is computed.

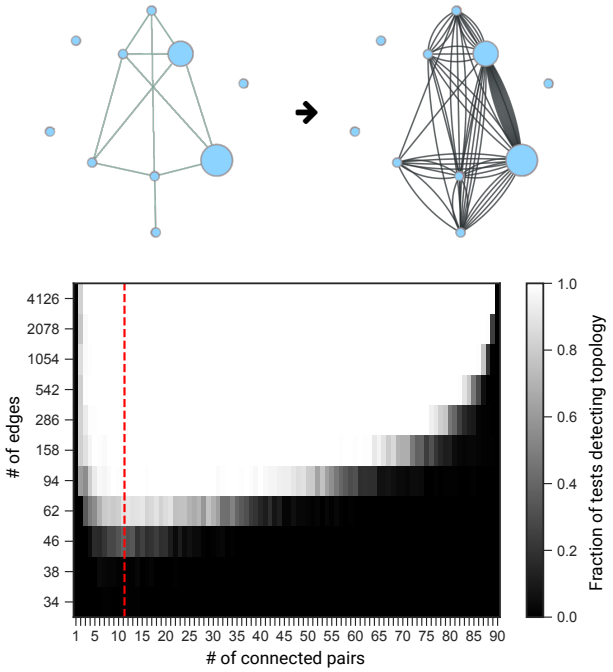


Figure 3.3: (Top) the topology on which interactions are allowed and a corresponding multi-edge network realisation. (Bottom) The fraction of 30 realisations of random interactions for which the null hypothesis of a mean-field model is rejected against the alternative that encodes the observes topology. Each realisation is generated by randomly drawing a number of m observations (y-axis) from a set of l possible interactions (x-axis).

We see a threshold $m = 45$, below which the topology is never detected. That threshold corresponds to the $l = 11$, or on average ≈ 4 interactions between each allowed pair[3]. At the limit $l = n(n-1)$, all pair-wise interactions are possible, meaning there is no more underlying topology behind the interactions. As expected, the network model is never selected at this limit. Similarly, the case when only one pair that can interact is also described by the configuration model. That is because there are two nodes with non-zero degrees and the observed network and the (one and only) realisation of the configuration model coincide. At first, for small l the nodes are only sparsely connected but with growing l, there are less and less nodes with degree zero. This sparsity combined with more nodes being connected by at least one edge means the topology becomes increasingly easy to detect. This trend is seen in Fig. 3.3 on the left side of the dashed red line.

That red line in the figure corresponds to *Molloy-Reed criterion*, which provides a threshold when a large connected component—meaning that there is a path between most nodes—is formed in a random network. This happens when the number of neighbours of neighbours of a random node is on average larger than the number of its own neighbours. It can be calculated for the simple model that we used to generate synthetic networks. With k'_i being the number of nodes with which node i can interact, the criterion writes in terms of the moments of the degree distribution as

$$\kappa = \frac{\langle k'^2_i \rangle}{\langle k'_i \rangle} > 2. \tag{3.14}$$

We can approximate the distribution of k' by writing it as an Erdös-Rényi model with a probability $p^{\mathrm{ER}} = l/(n^2 - n)$ to allow an arbitrary pair of nodes to interact. Then, the distribution of k' writes as a binomial, $k'_i \sim \mathcal{B}(n-1, \frac{l}{n(n-1)})$. Plugging the first and second moments of this distribution into the Eq. (3.14) leads to the condition for l:

$$l > l^{\mathrm{MR}} = \frac{n(n-1)}{n-2}. \tag{3.15}$$

For our example with $n = 10$ nodes we obtain $l^{\mathrm{MR}} = 11.25$. So, at the threshold of Molloy-Reed criterion a large connected component forms with the least number of edges between nodes. This creates a highly pronounced topology, which is the

[3]This number only refers to the studied example and cannot be taken as a general rule of thumb. It can vary depending on the network size and other properties such as how heterogeneous is the degree distribution.

easiest to detect. In the region $l > l^{\mathrm{MR}} = 11.25$ in Fig. 3.3 the trend changes for the detectability of the topology—one needs more observed edges m with growing l. In this region, the number of neighbours grows above one for an average node, coming closer to the configuration model. One can see the problem as a decrease of *disallowed pairs* and the hypothesis test between network and configuration models as a detection of these disallowed pairs. The fewer such pairs, the more interactions must be observed between the allowed ones to have a statistically detectable difference.

Zachary's Karate Club The above simple example allowed us to understand the behaviour of the likelihood-ratio test when comparing the most complex model to the simplest. Let us proceed to illustrating the procedure for the intermediate model with group structure.

We use the empirical based on frequencies of self-reported 231 interactions between 34 members of a university Karate club collected by Zachary [215], denoted in the following as $\hat{\mathbf{A}}$, shown in the top left panel of Fig. 3.4. For this networks with a known group structure, we can test whether this structure is sufficient to explain the observed interactions, thus using our model selection technique to test $\boldsymbol{\Omega}_g$ against the alternative hypothesis $\boldsymbol{\Omega}_n$. However, we first need to make sure that the configuration model $\boldsymbol{\Omega}_0 \equiv 1$ is not a sufficient model. According to the likelihood-ratio test $\Lambda(\boldsymbol{\Omega}_0, \boldsymbol{\Omega}_n)$ with 560 degrees of freedom [35], $\boldsymbol{\Omega}_0$ is rejected in favour of $\boldsymbol{\Omega}_n$ at significance level 0.05 (p-value $= 0.0052$). Similarly, $\boldsymbol{\Omega}_0$ against $\boldsymbol{\Omega}_g$ is rejected with p-value $= 2.32 \cdot 10^{-44}$ (likelihood ratio test with 1 degree of freedom). We then test $\boldsymbol{\Omega}_g$ against $\boldsymbol{\Omega}_n$. In this case we cannot reject the hypothesis $\boldsymbol{\Omega}_g$ (p-value $= 0.9996 > 0.05$, 559 degrees of freedom.) There is therefore no evidence for the (undirected, weighted) network hypothesis. However, a simple configuration model is not enough to describe the dataset, and a more complex group model is required. The ratio of odds between in-group and across-group interactions according to the fitted block matrix $\boldsymbol{\Omega}_g$ is $\hat{\omega} = c^{-1} = 10.53$. Hence, a node is approximately ten times more likely to connect to a node of the same group than to a node of the other group.

Let us now illustrate how Mahalanobis distance can be used to assess the goodness-of-fit of a model. Here we analyse two models, the configuration model $\boldsymbol{\Omega}_0 \equiv 1$ and the group model $\boldsymbol{\Omega}_g$ defined in Eq. (3.10). We test the hypothesis that $\boldsymbol{\Omega}_0$ is sufficient to describe the data by computing $\Pr\left[D_0^2(\mathbf{A}) \geq D_0^2(\hat{\mathbf{A}})\right]$ based on the distribution of Mahalanobis distances for random realisations $\mathbf{A}$ drawn from the ensemble. As expected, we obtain $p \approx 0$, meaning we can safely reject the hypothesis of $\boldsymbol{\Omega}_0$ (cf. red

histogram in Fig. 3.4). This outcome is in accordance with the likelihood-ratio test and can also be visually confirmed by comparing the empirical network in the top left of Fig. 3.4 to the random realisation of the (unbiased) ensemble shown in top middle panel of Fig. 3.4.

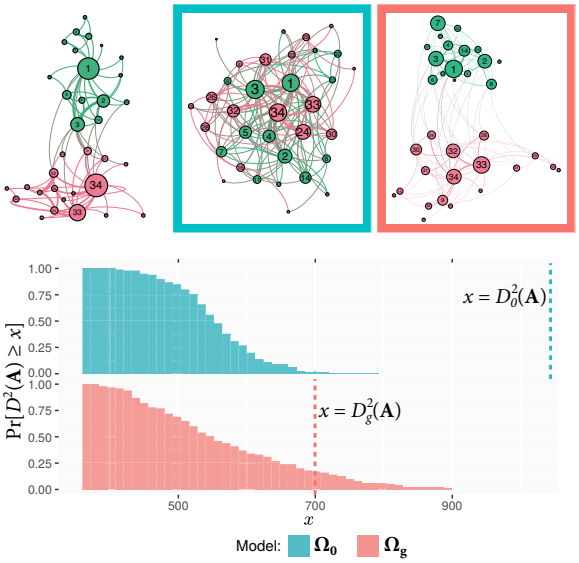


Figure 3.4: Using Mahalanobis distance to detect community structure in Karate club network. The top panel shows (left) the empirical Karate club network, (middle) a random realisation drawn from the unbiased hypergeometric ensemble, cf. Eq. (3.1), and (right) a random realisation drawn from a generalised hypergeometric ensemble with a block matrix $\mathbf{\Omega}_g$, cf. Eq. (3.5). The bottom panel shows the CCDF of Mahalanobis distances obtained for 5000 random realisations of (blue) the unbiased hypergeometric ensemble and (red) of the generalised hypergeometric ensemble with block matrix $\mathbf{\Omega}_g$. Dashed lines indicate the Mahalanobis distance for the empirical network in the two ensembles.

The second hypothesis states that the network topology is explained by the node degrees and the presence of two communities, where pairs of nodes within a community have higher propensities than nodes in different ones. Choosing c as the observed fraction of edges across communities—which is approximately equal the $c = \hat{\omega}^{-1} = 0.095$ identified by the likelihood ratio test—allows us to calculate the distribution of Mahalanobis distances for random realisations of the resulting statistical ensemble (cf. red histogram in Fig. 3.4). From this, we obtain $p = 0.158367$, which does not allow us to reject hypothesis of $\mathbf{\Omega}_g$. Again, this result is in accordance with the visually similarity between empirical Karate club network shown in the top left of Fig. 3.4 with

the random realisation generated from the group model shown in the top right of Fig. 3.4. The example shows that a generative model *only* accounting for heterogeneous node degrees and community structure is sufficient to statistically explain the observed network. Moreover, this highlights how the known functional form of distribution, expected values and covariance provided by our ensemble formulation provides a novel approach to (i) statistically test hypotheses in networks, and (ii) assess the significance of community structures.

3.3 Network reconstruction

A considerable scientific effort addresses the issue of network reconstruction [141, 190]. Specifically, one line of research aims to reconstruct the unknown edges in a network from aggregate statistics of nodes (e.g., from degrees) [134, 190]. Another line of research focuses on identifying the significant latent structure behind noisy pairwise interactions [46, 85, 142, 177, 184]. Most of the works in the second line follow a common approach—to identify the pairs of nodes that interact more often than expected at random. This approach is often called *inference of network backbone* [78].

In this section we show how our proposed generalised hypergeometric network ensembles can be used to infer the network backbone. We present two ways of achieving this goal. First, we show that the inferred edge propensities $\mathbf{\Omega}_n$ can provide a better proxy for the ground-truth topology underlying the observed pairwise interactions than the raw numbers of these interactions. Second, we follow the above-mentioned common approach and use our (unbiased) ensemble to identify the pairs of nodes that interact significantly more frequently than the model predicts.

3.3.1 Agent-based model for co-locations

In order to assess the goodness of network reconstruction methods, data with ground-truth information is needed. Obtaining such data of good quality is a challenge [148, 150]. To overcome this challenge, we use an agent-based model (ABM) [175]. These are computational models simulating actions of many elements, *agents*, and the interactions between them. The agents have internal properties described by a set of parameters. Hence, they are also described by *internal degrees of freedom*. An agent-based model defines a set of *microscopic rules* for the agents, which, together with the

internal properties of the agents, identify how they act and interact. ABMs are commonly employed to understand how *microscopic* interactions between many agents lead to the *macroscopic* properties of the studied system. They are especially useful when such *micro-macro* link is not a simple superposition of individual interactions, or cannot easily modelled analytically. In such cases the corresponding macroscopic properties are often called *emergent*.

In this section we utilise the ABM developed in [164]. In our model agents move on a two dimensional grid lattice, representing the mobility of humans [144, 168, 187, 196]. The agents also exhibit social behaviour by means of movements coordinated with their "friends". As a starting point for formulating out model, findings from three recent studies are used: Schneider *et al.* [168], Song *et al.* [187], and Toole *et al.* [196]. These findings can be summarised as follows.

- The mobility patterns of individual agents is time-dependent, meaning, in particular, that there are times of high activity (agent is "active" if it is moving) and low activity
- Each agent is characterised by a certain "home" location, at which its movements originate and end
- Agents show preference for certain locations by visiting them more often
- Pairs of agents are characterised by friendship ties
- Agents meet with friends at certain locations

Let us now formally define the model in brief based on the above statements (we refer to the thesis by Samarin for the full details [164]). We consider $n = 100$ agents on a square lattice of $L \times L$ cells with $L = 50$. Each agent i is assigned a home cell $h_i \in L \times L$ uniformly at random. An agent is characterised by an internal boolean state, which defines whether it is *active*. This state is by default zero, meaning that the agent is *inactive*. An agent turns active at a given time with probability proportional to a time-dependent global activity function $\alpha(t)$. This function is based on empirical observations in the (now defunct) location-sharing online platforms `Gowalla` and `Brightkite`[4] and exhibits a periodical daily pattern. It is higher during daytime and lower during night time.

[4] the data sets are available athttp://snap.stanford.edu/data/ (retrieved April 4, 2016)

Once an agent changes its state to active, it moves to a different cell. The new cell is chosen with a probability decreasing with the distance from the current cell according to power-law [25]. An agent stays in the new cell for certain time Δt (measured in seconds) drawn from a log-normal *waiting time distribution* $\Delta t \sim \ln \mathcal{N}(\mu, \sigma)$ with $\mu = 9.75$ and $\sigma = 1$:

$$\Pr(\Delta t) = \frac{1}{\sqrt{2\pi}\sigma\Delta t} e^{-\frac{(\log(\Delta t)-\mu)^2}{2\sigma^2}}. \tag{3.16}$$

After waiting for Δt in the visited cell, the activity state of the agent is reassessed. With a probability proportional to $A\alpha(t)$, $A = 15$, the agent remains active and visits another cell[5]. Otherwise, the agent turns inactive and return to the home cell.

The choice of the cell that an active agent visits is driven by one of two modes, which we call *regularity* and *exploration*. In the regularity mode the agent visits a cell that it previously visited and in the exploration mode it visits a previously unvisited cell. The probability of being in the exploration mode is determined by the number S of unique cells that the agent has previously visited as

$$\Pr(\text{explore}) = \rho S^{-\gamma}. \tag{3.17}$$

Otherwise, the agent follows the regularity mode and chooses to visit a cell from previously visited ones uniformly at random. According to Eq. (3.17), the probability to explore decreases with the number of unique cells previously visited by the agent, in accordance with the empirical findings in [187]. In the following the parameters in Eq. (3.17) are set to $\rho = 0.5$, $\gamma = 0.5$.

Last but not least, we define the rule for social behaviour of the agents. Underlying this behaviour is a friendship network $G^{fr} = (V, E^{fr})$ between agents. This network is generated by the configuration model with a degrees of nodes drawn according to power-law distribution $\Pr(k_i = k) \sim k^{-2.6}$, which is the best fit to both the empirical data sets `Brightkite` and `Gowalla`. The movements of agents' are influenced by the friendship network as follows. At the moment when an active agent moves to a cell, it decides whether it takes a friend with a probability β. If it decides to take a friend, one of its neighbours in the friendship network is chosen at uniformly at random and that friend agent moves to the same cell as the "inviting" agent. The state of the friend also changes to active, if it was not before the invitation, and it follows the rules described

[5] parameter A should not be confused with the adjacency matrix $\mathbf{A}$ with elements A_{ij}

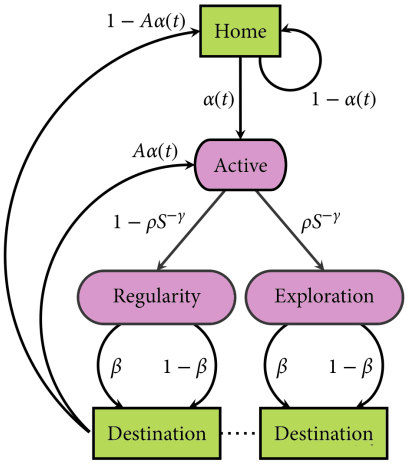


Figure 3.5: Schematic representation of the co-location ABM. Once an agents activates with probability $\alpha(t)$, it visits a new site with probability $\rho S^{-\gamma}$ or goes to a previously visited site. The agent moves together with a friend with probability β. Once at destination site, the agent returns home with probability $1 - A\alpha(t)$ or continues to another position otherwise. Figure from Samarin [164].

above. Figure 3.5 illustrates schematically the presented ABM.

Inferring friendship from co-location With the agent-based model defined, we can investigate how well the friendship ties between agents can be inferred from their co-locations, which is a simultaneous observation of two agents at the same location. The co-locations between agents form a multi-edge network $\hat{G} = (V, E)$ with adjacency matrix $\mathbf{A}$. The top panel of Fig. 3.6 shows the aggregate distribution of the logarithm of the number of co-locations between any two nodes conditional on the friendship relation between the two. As expected, the two distributions for friends and non-friends are better separated for the higher value of parameter β.

Given these distributions we can phrase the question of inferring the friendship relations as a statistical *classification problem*. For this problem we assume that friendship relations are not known, hence the classification is *unsupervised*. We employ *linear discriminant analysis (LDA)* [15] as the classification method to infer the two classes, friends and non-friends. LDA works as follows. It assumes that the two conditional probabilities $\Pr(\log A_{ij} \mid (i,j) \in E^{fr})$ and $\Pr(\log A_{ij} \mid (i,j) \notin E^{fr})$ are both normal with the same variance. With this assumption, it provides a weight for each value of $\log A_{ij}$

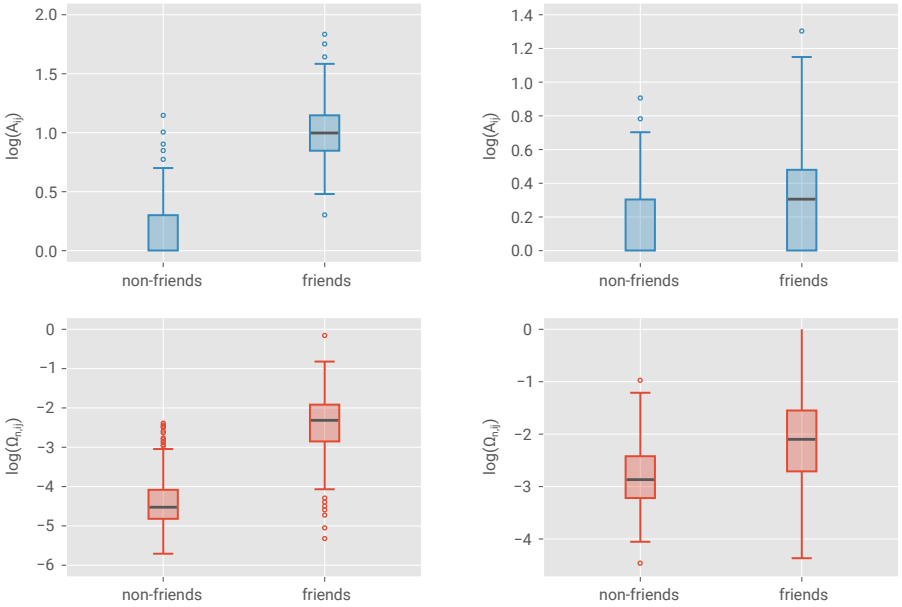


Figure 3.6: Distribution of (top) co-occurrence counts and (bottom) inferred edge propensities for friends and non-friends in the agent-based model with parameters (left) $\beta = 0.3$ and (right) $\beta = 0.05$.

to come from conditional distribution versus the other. Then a threshold value T on these weights maps to a threshold for $\log A_{ij}$, which separates the two classes.

In order to diagnose the quality of the classification, we use a standard measure—the *area under the receiver operating characteristic curve (AUROC)* [26]. It measures the area under the curve of the fraction of correctly identified friendships (*true positive rate*) against the fraction of incorrectly identified friendships (*false positive rate*) at various thresholds T. The closer the value of AUROC, the better is the classification. Figure 3.7 shows in blue the outcome of LDA based on the two cases shown in the top panel of Fig. 3.6, for $\beta = 0.3$ and $\beta = 0.05$. In addition to the co-locations counted at the (discrete) time steps of the ABM simulation, the outcome based on counts within time windows of different duration are shown (i.e., if time difference for two agents to visit the same cell is smaller than the time window, a co-location is registered). As expected, friendship relations are almost perfectly detected for stronger social behaviour ($\beta = 0.3$), i.e., when a considerable fraction of agent movements are accompanied by friends.

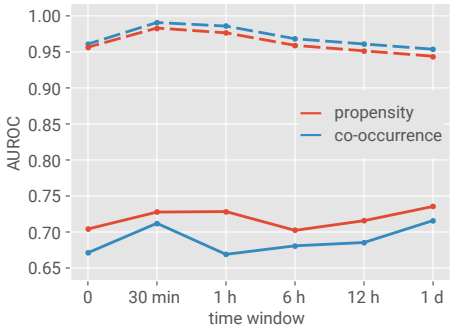


Figure 3.7: The AUROC for the friendship classification based on co-location counts and the inferred edge propensities. Dashed lines correspond to $\beta = 0.3$, solid lines correspond to $\beta = 0.05$.

Let us now fit the network model $\mathbf{\Omega}_n$ of generalised hypergeometric network ensemble to the multi-edge network of co-locations $\hat{G}$ and repeat the classification procedure based on $\mathbf{\Omega}_n$ instead of raw co-location counts. The bottom panel of Fig. 3.6 shows the distribution of $\log \Omega_{n,ij}$ conditional on the friendship edge between i and j. The red curves in Fig. 3.7 show the corresponding AUROC values. We see that for the (almost trivial) case of $\beta = 0.3$ the raw count of co-locations slightly outperform the fitted propensities for inferring the underlying friendship relations. However, when the fraction of friendship-driven co-locations is lower, $\beta = 0.05$, the classification is more difficult (smaller AUROC values) and the inferred propensities considerably outperfom the raw co-location counts.

3.3.2 Inference of social network from interaction data

Above we presented a propensity-based network reconstruction method on the example of data generated by an ABM. In the following, we demonstrate how our ensemble can be employed to infer the significant connected pairs following the common approach of backbone inference. We use the "*Reality Mining*" data set that captures time-stamped proximities between students and faculty at MIT [56] recorded via smart devices. We denote the weighted adjacency matrix capturing observed pairwise interactions as $\hat{\mathbf{A}}$. For a given significance threshold α, we then identify significant connected pairs by filtering matrix $\hat{\mathbf{A}}$ by a threshold $\Pr(A_{ij} \leq \hat{A}_{ij}) > 1 - \alpha$ based on Eq. 3.6. This can be seen as assigning p-values to pairs (i, j), obtaining a *high-pass* noise filter for entries in the adjacency matrix.

To illustrate our approach, Fig. 3.8(a) shows the entries of the (original) adjacency matrix $\mathbf{A}$ for *Reality Mining* data. The high-pass noise filter resulting from our methodology (using $\alpha = 0.01$) is shown in Figure 3.8(b), where black entries correspond to pairs of nodes that interacted at least once but which are identified as non-significant. The application of this filter to the original matrix yields the noise-filtered matrix shown in Fig. 3.8(c). While in the original network there are $721,889$ observed multi-edges amounting to $2,952$ distinct connected pairs, after filtering there are 626 (21.2%) significant connected pairs left ($617,069$ multi-edges, 85.5% of the original). We validate the benefit of filtering the original interactions in *Reality Mining* data by comparing the output of a standard community detection algorithm—the degree-corrected block model [151]—in (i) the original, unfiltered graph shown in Fig. 3.8(d), and (ii) the filtered, backbone network shown in Fig 3.8(f). Using known classes of students and affiliations of staff members as ground truth allows us to compare the quality of the community detection. Figure 3.8(e) shows the set overlaps between the ground truth labels (middle column) and detected partitions in the unfiltered (left column) and filtered graph (right column). Due to the high number of non-significant connected pairs in the unfiltered graph, the algorithm only detects three partitions, each spanning multiple labs and classes. In contrast, applying the algorithm to the filtered graph yields six partitions that better capture the ground truth lab and class structure. As expected, detected partitions do not perfectly correspond to the ground truth, since labs and classes are likely not the only driving force behind observed proximities.

A major advantage of generalised hypergeometric network ensemble against other backbone inference methods is that, by specifying a non-uniform matrix $\mathbf{\Omega}$, we can additionally encode known factors that influence the occurrence of interactions between nodes, while still obtaining an analytically tractable ensemble. In our second illustrative example, we use this to encode the known structure of two separate Karate classes in the Zachary's data set. These two classes naturally influence the frequency of encounters between actors beyond what would be expected at random according to the configuration model. As before, we incorporate this prior knowledge via a block matrix $\mathbf{\Omega}_g$ that assigns higher dyadic propensities to pairs of actors in the same class (cf. Eq. (3.10)). This establishes a statistical baseline accounting both (i) for combinatorial effects due to heterogeneous node degrees, and (ii) the known group structure in the data. Using a significance threshold of $\alpha = 0.01$, this yields the result that only 8 out of 78 interacting pairs are significant ($\sim 90\%$) of 231 observed multi-edges are filtered out, as shown in Fig. 3.9. In other words, taking into account the

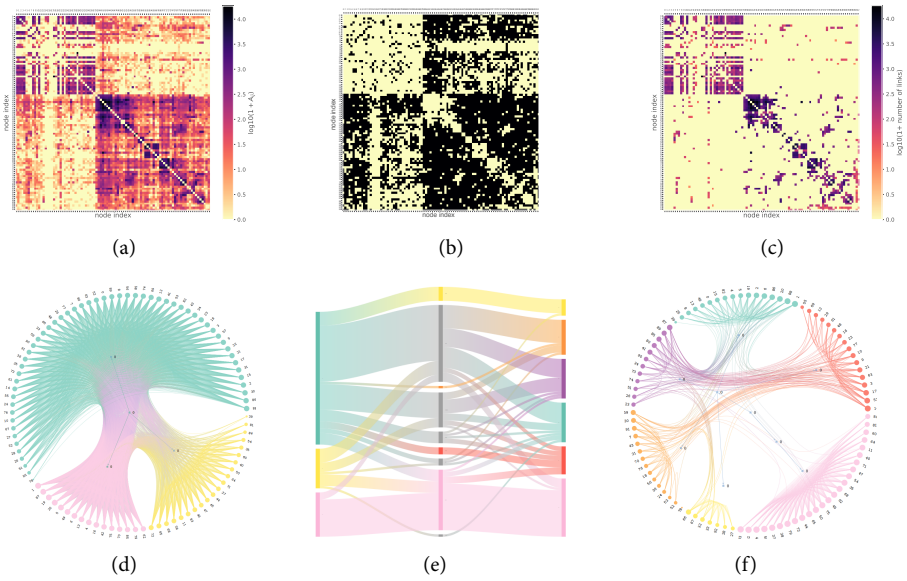


Figure 3.8: Illustration of our approach in the *Reality Mining* data set capturing proximity of students and staff at MIT campus. For the observed adjacency matrix (a) and a given significance threshold, our framework allows to establish a high-pass noise filter matrix (b), which can be used to obtain a filtered adjacency matrix containing only significant connected pairs (c). A visual comparison of the output of a community detection algorithm on the unfiltered (d) and filtered (f) graphs shows that detected partitions in the filtered one better correspond to ground truth lab affiliations and classes (e).

partitioning of members in two classes almost all encounters between club members can simply be explained by random effects. This is in strong agreement with the result in Section 3.2, where we found that the group model is sufficient to statistically explain the observed network.

3.4 Conclusion

In this chapter we introduced *generalised hypergeometric ensembles*, a broad class of statistical ensembles that allows to encode a wide-range of topological patterns. Unlike similar approaches, it provides analytical expressions for important statistical quantities for nodes and edges like expected values and covariance. Through this, the ensembles introduced in this chapter provide broad perspectives for the analysis of

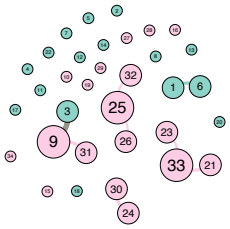


Figure 3.9: The filtered network for the Zachary's data set, capturing encounters between members of a Karate club. Most of the observed encounters can be explained by random effects resulting from the club members' separation into two classes.

complex networks, with applications in pattern recognition, hypothesis testing and statistical inference. This work contributes to the fundamentals of network analysis, with applications in the interdisciplinary study of complex systems in physics, biology, and (computational) social science. More specifically, the key contributions of this chapter are:

1 We introduced a broad class of *statistical ensembles*, i.e., probability spaces of weighted network topologies subject to a tunable set of constraints. Highlighting previously unknown relations between random graph theory and the multivariate Wallenius' non-central hypergeometric distribution, we established the analytical tractability of this class of ensembles. We further showed that it can be viewed as a natural generalisation of the frequently used configuration model for directed and undirected networks with a fixed sequence of degrees.

2 Building on this theoretical foundation, we developed a method to formalise and test hypotheses about the mechanisms driving *repeated interactions* between the elements in a complex system. It particularly allows to test whether a network model is justified or not, thus contributing to the the body of research on network inference. Apart from that, it further enables to test this *network hypothesis* against simpler, alternative models that can explain the data. Validating our method in synthetically generated data where the process that generates repeated interactions is known, we show that our framework reliably selects the correct model even when data are scarce.

3 We then apply our method to empirical data sets on social interactions, which are frequently used as examples for social networks. Remarkably, our analysis reveals that a network abstraction is not always justified. Instead, our framework highlights that the (known) group structures can be sufficient to

explain both the frequency and the topology of observed interactions. This ability to statistically investigate group structures in networks makes the method applicable in the field of community detection. Although, we did not study the applicability, performance and the relations of our method to other methods for community detection, we can already mention some such relations already now. For instance, testing a group structure as we have shown is quite similar in concept to the degree-corrected stochastic block modelling approach [100]. Other methods that implicitly or explicitly encode metadata about nodes into propensities are interesting to compare with our method [64, 137].

4 Finally, we showed how our ensemble can be used to infer the backbone of the network from the raw noisy data on interactions. We explored two ways of doing so. In one case we inferred the weighted *degree-corrected* network structure by means of fitting edge propensities. We demonstrated on the example of an agent-based model how this approach can outperform the raw counts of interactions in recovering social ties between interacting agents. In the second case, we used our proposed network ensemble as a statistical baseline (null model) and filtered the interactions that happen significantly more often than expected at random. In this line of application we did not present a comparison of our method to existing methods of backbone inference. This is an important task and is left for future research.

Our work advances our ability to decide how to model, interpret, and analyse relational data on repeated interactions. Highlighting challenges in the process of turning data on *interactions* into *network models*, we argue that it has major consequences for data-driven studies of social, biological and technical systems.

Chapter 4

Significant deviations in network topology

Summary

Social scientists have long been interested in signed social network, as positive human relations, such as friendship, are inevitably accompanied by negative interactions, such as animosity. However, there is currently a lack of data on negative interactions, and consequently, on signed social networks. We propose a procedure to infer signed relations from unsigned networks on repeated interactions. This procedure builds on a new statistical measure of deviations, which is valid for a broad range of distributions, including discrete, bound and skewed ones. Such are the distributions describing the null models of networks, to which we compare the observed repeated interactions, in order to infer the signed relations. We show how our method works on examples of both synthetically generated networks and empirical networks. As a validation of our method, we are able to reproduce and extend previous results for the cultural dynamics in the Eurovision song contest.

Based on Nanumyan V. "Measuring significant deviations in network topology", *working paper*.

More than a hundred years ago sociologist Georg Simmel criticised his colleagues for focusing too much on positive social relations, when, as he put it, society is built on both "harmony and disharmony, association and competition, affection and resentment" [179]. Today, network science seems to be in a similar state— most studies focus exclusively on positive relations and interactions, such as friendship, follower and collaboration relations, likes, re-tweets etc. [111]. Yet, the reason for this is more practical than philosophical. The overwhelming majority of online platforms with a social interaction component *by design* do not offer their users tools to express negative relations. Likewise, most people in everyday life are hesitant to publicly announce their enmity to somebody. And even in a field like international relations, where we have some openly hostile relations (e.g., wars, sanctions), we can safely assume that there is a much greater number of more subtle negative relations, which cannot be detected and quantified easily. Having networks with both positive and negative relations, however, is crucial for testing sociological theories such as *structural balance theory* mentioned in Chapter 1 [34, 50]. It must be noted, though, that Jacob Moreno constructed and visualised the first signed networks as early as 1934 in his work *Who Shall Survive* [130]. He called these *sociograms*, one of which we show in Fig. 4.1.

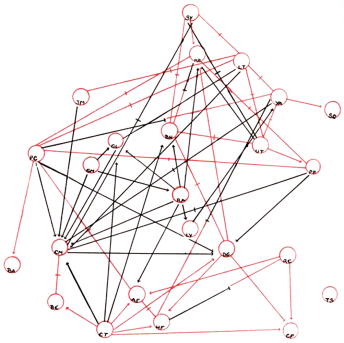


Figure 4.1: A signed *sociogram* by Jacob Moreno, reprinted from page 518 of [130].

Moreno introduced his sociograms so early that he not only pioneered the study of signed networks, but of social networks in general. The fact that these earliest representations of social networks were already signed is an indication that the current sparsity of research on signed networks is not because of their irrelevance, but due to the aforementioned lack of data. How can we solve this problem of missing data on negative relations? Previous studies have applied various methods to ascribe a positive or negative sign to existing observed relations [110, 118, 213]. However, they assume that negative relations are explicitly present in the network data, and it is merely unknown *which* edges are positive or negative. Instead of simply labelling the unsigned edges, we hereby propose a method to *infer signed relations from the counts of repeated interactions*. Our method is based on the conjecture that a large amount of interactions between a pair of social agents is a signal of a positive relation, and, vice versa, an unexpected lack of interactions is a signal of a negative relation.

Of course, not every non-existing positive interaction can be translated into a negative one—in large-scale societies, individuals only interact with a small subset of others [192]. However, this does not mean that they have a negative relation with all other individuals. What is needed for judging whether a lack of interactions is expected or is a signal of a negative relation, is a *null model* that, based on reasonable assumptions, tells us between which nodes in the network we should expect positive interactions, and specifically *how many* positive interactions. As outlined in Section 1.2, by comparing the probability distribution generated by the null model to the actual observed count of interactions, we can then determine whether two given nodes interacted significantly more than expected (a positive relation), as expected (neutral), or less than expected (the sought-after negative relation). As a result, we receive a *signed, weighted network* that will allow us to see both sides of society, just as Georg Simmel intended.

In Chapter 3, we have already introduced a flexible network ensemble that is able to encode various topological structures. These ensembles can serve as null models based on which we can transform unsigned interaction counts into signed relations. Such a null model provides the marginal probability distributions for the interaction counts between any pair of nodes, cf. Eq. (3.6). What we miss is a method that outputs a signed value from a comparison of the observed number of interactions between two nodes to the corresponding marginal distribution in the ensemble. Effectively, this method will measure the deviation of an observation in the distribution. Ideally, the resulting signed value will have a statistical interpretation, e.g., the significance of the deviation.

To the best of our knowledge and to our surprise, we find that no existing statistical measure fits our purpose. In the following, we formulate a new statistical measure of deviation, preceded by a discussion of related measures and the reasons why they do not satisfy our needs.

4.1 Signed measure of deviation

There is a widely used statistical measure of deviation—the *standard score*, also known as the *Z-score*, which shows how many standard deviations away the value of an observation is from the mean of a normal distribution [188]. For a random variable x,

the Z-score is

$$z = \frac{x - \mu}{\sigma}, \tag{4.1}$$

where μ is the mean of the distribution and σ is the standard deviation. So, the Z-score is positive if x is above the mean μ and is negative otherwise. This score is often used in statistical testing when the random variable can be approximated by normal distribution. In such cases, the value of the Z-score maps to a significance level. For instance, $z = 2.326$ corresponds to 0.01 significance level.

Another statistical method related to our problem is the *mean normalisation*,

$$x' = \frac{x - \mu}{x_{\max} - x_{\min}}, \tag{4.2}$$

where $x_{\max}$ is the maximum and $x_{\min}$ is the minimum value of the random variable x. Mean normalisation is a case of *feature scaling*, which is used in data processing to standardise the range of the data. It is commonly performed as a pre-processing step for discriminative statistics or machine learning techniques. Like the Z-score, mean normalisation produces a signed value that measures the deviation from the mean. In this case, however, this deviation is measured in relation to the full range of values of the random variable. The above two measures are parametric. While the mean normalisation has one parameter less than the Z-score and, unlike Z-score, does not assume a specific distribution, it does not provide a statistical interpretation.

What we need—instead of a measure that relies on asymptotic theory—is an *exact measure* [45], meaning that all the assumptions behind the statistical interpretation of the measure are met. A major reason for this is that we must be able to compare the inferred signed relations between different pairs of nodes, even if these pairs differ strongly in their centrality in the network. For instance, in a network with a broad degree distribution, the value of the signed relation between the two nodes with the highest degrees must have the same statistical interpretation as for the two nodes with the lowest degrees. In this case, a measure relying on asymptotic theory would fail, because the two marginal distributions for the two different pairs differ drastically (cf. Eq. (3.6))—while for the pair of high-degree nodes the distribution may be well approximated, e.g., by a continuous distribution, for the low-degree nodes this distribution can take very few values. Hence, the discreteness of the distributions must be accounted for in an exact manner. The same argumentation holds when we want to compare two pairs of nodes that are described by high and low edge propensities. In that case, the skewness of the two distributions is very different. Thus, an approximate

measure would disregard this difference.

More generally, the following properties of the marginal distributions of the ensemble, shown in Eq. (3.6), make the above-mentioned measures unsuitable: (i) it is bounded, meaning there is a finite range of values that the random variable can take, (ii) it is skewed, meaning that it is not symmetric around the mean, (iii) it is discrete as it can take only integer values. A possible nonparametric solution that would allow comparing pairs of nodes with strongly dissimilar properties is to define an exact measure based on the percentiles of the underlying distribution, i.e., the exact shape of the distribution. Before we introduce a simple measure based on the cumulative distribution function, $F(x) = \Pr(X \leq x)$, let us formulate the conditions that we want such a measure to satisfy.

Conditions to fulfill For a discrete random variable $X \in \mathbb{N}$ or a continuous $X \in \mathbb{R}$ distributed according to a known probability distribution, we want the measure of deviation $\Phi(x)$ to fulfil the following conditions:

1 The measure must be unbiased, meaning the expectation of the measure is zero, $\mathrm{E}[\Phi] = 0$

2 The measure must have a fixed range, such as $\Phi : x \to [-1, 1]$, with the boundary values of Φ meaning the most extreme deviations from the distribution

3 There is a central value $\bar{x}$ for which $\Phi(\bar{x} - a) < 0$ and $\Phi(\bar{x} + a) > 0$ for any $a > 0, a \in X$.

4 The absolute value of Φ is symmetric if the distribution of X is symmetric, i.e., there is a $\tilde{x}$, such that for any $x \in X$

$$|\Phi(\tilde{x} - x)| = |\Phi(\tilde{x} + x)|,$$

if $\Pr(\tilde{x} - x) = \Pr(\tilde{x} + x)$

A measure that satisfies the above conditions is illustrated in Fig. 1.3 by means of the colour fill under the distribution function, with the lightest yellow representing the zero value of the measure, darker shades of green representing higher positive values and the darker shades of red representing lower negative values.

4.1.1 Continuous random variable

Let us start with the case of a continuous distribution. To define the measure Φ, we need to identify two components: the measure of central tendency and the functional form that maps the observation to the signed value. As mentioned above, our objective is to define a nonparametric measure that can be applied to sample distributions without making strong assumptions about the underlying distribution, such as the existence and the form of the moments. As it is shown below, the (sample) median as the measure of central tendency and the (empirical) cumulative distribution function serve this purpose.

We define $\Phi(x)$ as the probability to randomly obtain any value X between the median m_X of the probability distribution of X and the observed value x with the sign depending on whether x is smaller or larger than the median:

$$\Phi(x) := \begin{cases} -2\Pr(x < X \leq m_X), & \text{for } x < m_X, \\ 0, & \text{for } x = m_X, \\ 2\Pr(m_X < X \leq x), & \text{for } x > m_X, \end{cases} \tag{4.3}$$

which simplifies to

$$\Phi(x) = 2\Pr(X \leq x) - 1 \tag{4.4}$$

All the aforementioned conditions are straightforwardly satisfied:

- $\mathrm{E}[\Phi] = 2\,\mathrm{E}[\Pr(X \leq x)] - 1 = 0$, as the the cumulative distribution function is uniformly distributed in $[0, 1]$ with $\mathrm{E}[\Pr(X \leq x)] = 0.5$
- The values of Φ range continuously from $\Phi(\min X) = -1$ to $\Phi(\max X) = 1$
- $\Phi(m_X - a) < 0$ and $\Phi(m_X + a) = 0$ for $a > 0$ and $\Phi(m_X) = 0$
- If the distribution is symmetric, the median—which coincides with the mean—is the symmetry point and it directly follows from Eq. (4.3) $\Phi(x)$ is also symmetric.

4.1.2 Discrete random variable

The signed measure of deviation formulated in Eq. (4.4) does not satisfy the conditions we have set for a discrete random variable $X \in \mathbb{N}$. Let us show this with an example for

conditions 1 and 4. To see that condition 1 does not hold, we consider the binomial distribution $B(x, p)$ of $x = 1$ trial with a low probability of success $p = 0.1$. Then, the expectation of Φ according to Eq. (4.4) is

$$\begin{aligned} \mathrm{E}[\Phi \mid B(1, 0.1)] &= \sum_{n=0}^{1} \Phi(n) \binom{x}{n} p^n (1-p)^{x-n} \\ &= 0.9\Phi(0) + 0.1\Phi(1) \\ &= 0.9 \cdot 0.8 + 0.1 \cdot 1 \\ &= 0.82 \neq 0 \end{aligned} \tag{4.5}$$

Next, consider the case of binomial distribution with $p = 0.5$ and $x = 2$ trials, $B(2, 0.5)$, which is symmetric around $x = 1$: $\Pr(1) = 0.5$ and $\Pr(0) = \Pr(2) = 0.25$. Condition 4 does not hold for this example, because according to Eq. (4.4), $\Phi(0) = -0.5$ and $\Phi(2) = 1 \neq \Phi(0)$.

In order to fulfil both conditions for the discrete case, we redefine Φ for the discrete variable as follows:

$$\Phi(x) = \Pr(X < x) - \Pr(X > x). \tag{4.6}$$

Note that this definition is equivalent to Eq. (4.4) in the continuous case when the random variable X has a probability density function defined, because then $\Pr(X < x) = \Pr(X \leq x)$ and $\Pr(X > x) = 1 - \Pr(X \leq x)$. It can be proven that the measure Φ defined in Eq. (4.6) now satisfies conditions 1 and 4 for a discrete variable (which we will show by means of example in the following). Like in the case of continuous variable, the median m_X is the value required by condition 3 at which the measure changes its sign.

The range of values of Φ Condition 3 is not satisfied if the random variable is not continuous. Specifically, it does not hold if the minimum or maximum values of the random variable have a finite probability:

$$\Phi(x_{\min}) = \Pr(x_{\min}) - 1 = \min(\Phi(X)) \tag{4.7}$$

$$\Phi(x_{\max}) = 1 - \Pr(x_{\max}) = \max(\Phi(X)) \tag{4.8}$$

This behaviour, however, is desirable (as we explain below) and instead of further refining the measure Φ, we relax the condition 2 to $\Phi(x) \subset [-1, 1]$. With this, the interpretation of the value of $\Phi(x)$ is the same independent of the distribution from

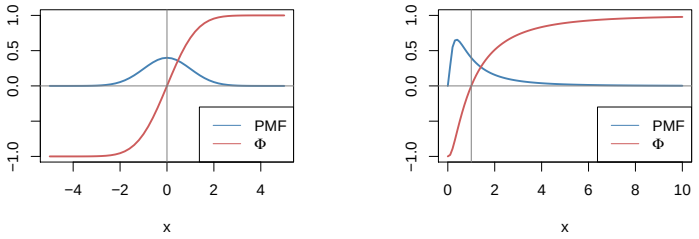
which x is drawn: the higher the absolute value of the measure, the more unlikely it is for the random variable to be so far in the tail of the distribution. So, if a boundary value x_b of the random variable, $x_{\min}$ or $x_{\max}$, has a finite probability, then it is not so unlikely that it is observed, limiting the corresponding value of $|\Phi(x_b)| < 1$. At the same time, the sign of the measure indicates in which tail of the distribution the observation is—left for a negative value and right for a positive.

The distribution-independent interpretation of the measure allows comparing observations that originate from different distributions, which has profound value for our goal of inferring social relations from observing interactions. For the integer-valued marginal distributions of network ensembles that describe repeated interactions between nodes, zero is the minimum of the random variable describing the count of interactions, and the extent to which the zero value is an under-representation will depend on $\Pr(0) = p_0$. Coming back to the example of two pairs of nodes, we see that according to the configuration model, the probability to observe no interaction between the high-degree nodes is very low, $p_0 \approx 0$, leading to a low value of $\Phi(0) = p_0 - 1 \sim -1$, while the probability of no interactions between low-degree nodes will be high, $p_0 > 0$, leading to a moderate value of $\Phi(0) > -1$ (cf. Eq. (3.2) and Eq. (4.7)). Hence, not observing any interactions between high-degree nodes is a severe under-representation, while it is less significant in the case of low-degree nodes.

The definitions provided in Eqs. (4.4) and (4.6) allow to interpret the values of the measured deviations in terms of statistical significance. In the continuous case, we can say that an observation x is significantly larger than the median of a given distribution at a significance level α if $\Phi(x) > 1 - 2\alpha$, which directly follows from Eq. (4.4). Similarly, the observation is significantly smaller than the median at a significance level α if $\Phi(x) < -1 + 2\alpha$. The above conditions are not exactly true for the case of a discrete variable (Eq. (4.6)), however the general interpretation of the signed measure is the same. Hence, in our applications we will threshold the measure Φ at a given level α as simply $|\Phi| > 1 - \alpha$ but we will not call this threshold a "significance level" (which has a precise meaning in statistics).

4.1.3 Examples of common distributions

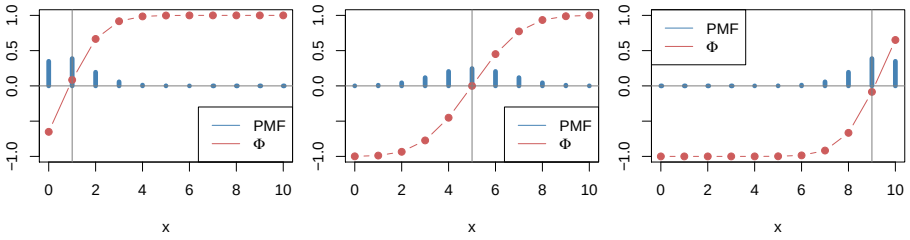
Before applying the measure Φ to networks of repeated interactions, let us first investigate its behaviour in the case of continuous and discrete, symmetric and asymmetric common distributions. Figure 4.2 shows the function $\Phi(x)$ for two different con-



(a) Normal with $\mu = 0,\ \sigma = 1$

(b) Log-normal with $\mu = 0,\ \sigma = 1$

Figure 4.2: The measure Φ applied to Normal and Log-normal random variables. The vertical line in each plot indicates the median of the distribution.



(a) $n = 10,\ p = 0.1$

(b) $n = 10,\ p = 0.5$

(c) $n = 10,\ p = 0.9$

Figure 4.3: The measure Φ applied to binomial random variable. The vertical line in each plot indicates the median of the distribution.

tinuous distributions—symmetric Normal and positive-valued, skewed Log-normal. As shown in Fig. 4.2(a), $\Phi(x)$ is a symmetric function for the symmetric Normal distribution $\mathcal{N}(0, 1)$. For both distributions the range of values of $\Phi(x)$ is $[-1, 1]$ and equals to zero at the median m_X.

Figure 4.3 illustrates the discrete case by the example of binomial distribution. The two differences from the continuous cases are the following. First, the range of $\Phi(x)$ is not the whole $[-1, 1]$ interval if the minimum or maximum of the random variable has a finite probability, shown respectively in Fig. 4.3(a) and Fig. 4.3(c). Second, at the median, the value $\Phi(m_X)$ is not necessarily zero, however the two neighbouring values have a different sign: $\Phi(m_X - 1) < 0$ and $\Phi(m_X + 1) > 0$.

4.2 Signed network from interaction counts

In Chapter 3, we introduced a novel statistical ensemble of multi-edge networks. There are various other methods for fitting random network models to data on repeated interactions. These include *Stochastic Block Models* [92, 100] for describing community structures and, for more general purposes, the thoroughly studied *Exponential Random Graph Models (ERGM)* [163, 186]. All these are ensemble-based methods, meaning they are described by a multivariate distribution, based on which we can apply the measure Φ to find how much the observed interactions between individual pairs deviate from the model.

Depending on how the network model is fitted and selected, the signed relations produced by the measure Φ can have different applications. For instance, they can be used for outlier detection: a network model can be a good fit for the network as a whole, except for some outlier pairs of nodes, which will be indicated by an extreme value $|\Phi| = 1$. Once the outliers are detected, the network model can be further improved by disregarding these.

The most important application of the method, however, is to use a network ensemble as a baseline, a null model, and to infer a signed network from it. Then, the edges in this signed network represent deviations of the observed interactions from the null model, highlighting the fact that there have been aspects of the network formation behind the observed network that are not included in the null model. We will apply the method in this manner to citation networks between authors in Chapter 5, outlining a procedure to study the extent to which citations between authors go beyond purely scientific reasons.

The signed relations inferred from a combinatorial null model (e.g., configuration model) also allows for inferring the backbone of a network. In Section 3.3, we followed a common procedure for backbone inference based on percentiles. That procedure can be improved using our signed measure. Assume there are $\hat{A}_{ij} = 10$ interactions observed between a given pair of nodes i and j, and the corresponding marginal distribution of the ensemble resembles the distribution shown in Fig. 4.3(c). The percentile-based backbone inference would select this pair as part of the backbone for any threshold, as $\Pr(x \leq \hat{A}_{ij}) = 1$. However, we see that the observed 10 interactions are not at all indicating an over-representation (the criterion for such backbone inference techniques), which is quantified by a $\Phi(\hat{A}_{ij}) = 0.651$. Hence, the pair will not be included in the backbone at any reasonable threshold applied to our

signed network.

4.2.1 Synthetic networks

In order to better understand the effect of applying the measure Φ defined in Eq. (4.6) to networks of repeated interactions, let us start with simple synthetic examples of random multi-edge networks.

Erdös-Rényi network We generate a multi-edge network according to the $G(n, m)$ variant of the Erdös-Rényi model with n nodes and m total edges such that each pair of nodes has the same probability to receive an edge, $p_{ij} = 1/(n^2 - 1)$ for any nodes i and j. Without affecting the further discussion, we do not allow self-edges, so $p_{ii} = 0$. This model is described by the multinomial distribution, i.e., the integer-valued symmetric adjacency matrix $\hat{\mathbf{A}}$ that encodes the multiplicity of each link, $\hat{A}_{ij} = \hat{A}_{ji} \in N$ and $\hat{A}_{ii} = 0$, has the probability

$$\Pr(\hat{\mathbf{A}} = \mathbf{x}) = \frac{m!}{\prod_{i\in\{1,n\}} \prod_{j>i} x_{ij}!} \left(\frac{2}{n(n-1)} \right)^m \tag{4.9}$$

Figures 4.4(a) and 4.4(b) show the adjacency matrices for two realisations of the model for $n = 20$ nodes and, correspondingly, $m = 100$ and $m = 1000$ edges among $n(n-1)/2 = 190$ pairs of nodes. We can now apply the measure Φ to obtain the signed matrix $\Phi_{ij} := \Phi(A_{ij})$ using the marginal distributions of the unbiased hypergeometric ensemble (the configuration model) given by Eq. (3.2) for each pair (i, j). Doing so will show how much the number of interactions between each pair of nodes deviates from the configuration model. As in Chapter 3, we build the ensemble based on the degrees of nodes, $k_i = \sum_j A_{ij}$.

Let us now compare the resulting matrices $\mathbf{\Phi}$, which are shown for the two networks in Figs. 4.4(c) and 4.4(d). Visually both matrices are random, as expected from the construction of the networks. One can see, however, that the matrix corresponding to the network with more edges (Fig. 4.4(d)) has generally brighter coloured elements meaning that there are stronger deviations in this network. This observation is confirmed by the plotting of histograms of the values Φ_{ij} for all $n(n-1)$ pairs of nodes, shown in Figs. 4.4(g) and 4.4(h)[1]. These histograms show that for the first network

[1] One could think that there are as many values in the histogram as $\max A_{ij}$, which is not the case.

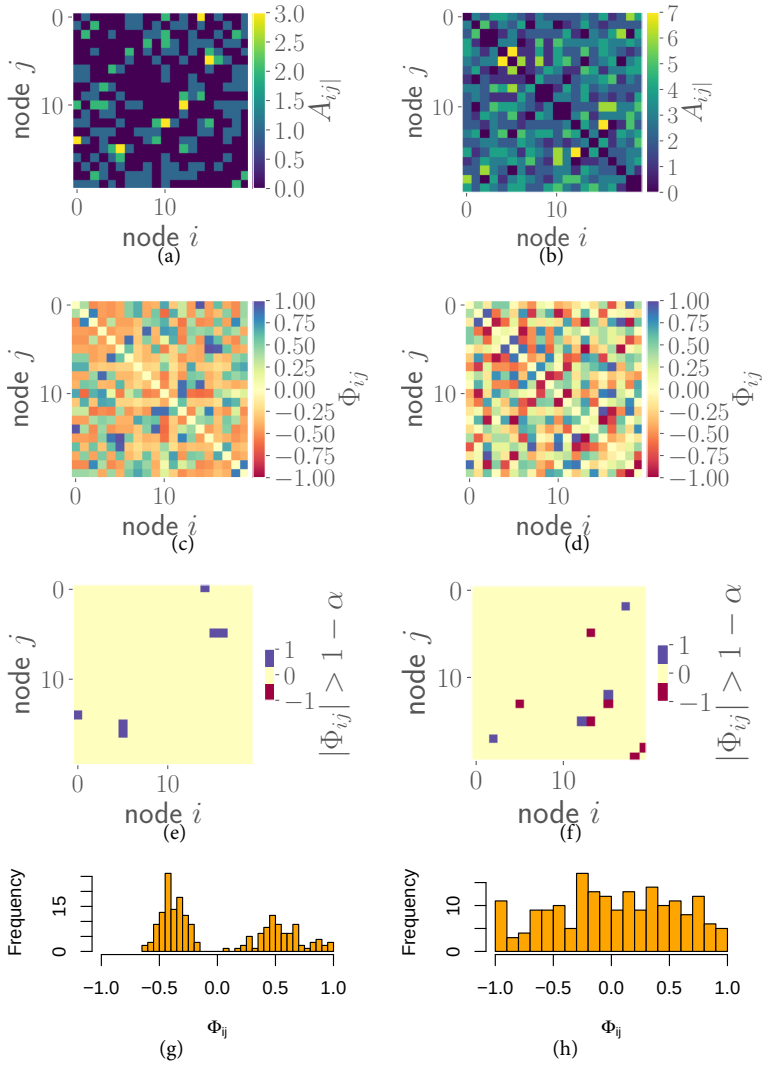


Figure 4.4: Two random networks generated according to Eq. (4.9) with (left column) $m = 100$ and (right column) $m = 500$ edges. The first row shows the networks by means of adjacency matrices $\mathbf{A}$, the second row shows the matrix of deviations Φ_{ij} computed based on the configuration model. The third row shows the histograms of values of Φ_{ij} for all $n(n-1)$ pairs of nodes. The forth row shows the node pairs with significantly under-represented ($\Phi_{ij} < -1 + \alpha$) and over-represented ($\Phi_{ij} > 1 - \alpha$) edges at level $\alpha = 0.05$.

with fewer edges, the distribution of Φ_{ij} values is bimodal—meaning that there are two pronounced peaks in the distribution—while it is much closer to a uniform distribution for the second network with more edges. Moreover, we see that for the first network, the part of the distribution for negative values is narrower (with a higher peak) than the part for positive values. This observation can be explained as follows. With $m = 100$ edges among $n = 20$ nodes there is on average $2M/(n^2-n) = 0.526$ edge per pair of nodes, meaning that the network is relatively sparse with roughly half of the pairs not having any edge. For all these pairs (i, j), the signed relation will be negative, $\Phi_{ij} < 0$. The absolute value of the relation is determined by the degrees of the two nodes i and j. In our randomly generated network, the degrees of nodes are narrowly distributed[2] around the average $\langle k \rangle = m/n = 5$, meaning that the negative values of Φ_{ij} for disconnected nodes will also be narrowly distributed around some average negative value (the left peak in Fig. 4.4(g)). So, for our relatively sparse network, the signed matrix $\mathbf{\Phi}$ visualises as broadly distributed positive values embedded in a mildly negative "background" corresponding to the disconnected pairs of nodes. This highlights an important aspect of inferring signed relations from repeated interactions: strongly negative relations inferred from a configuration model can be observed only if two high-degree nodes have only few or no interactions with each other. In the context of social interactions, this translates to the following. One can identify that two social agents are avoiding each other, only if they both interact abundantly with others, but not with each other. This strong dependence on total interactions is mitigated if the underlying null model encodes heterogeneous propensities.

Returning to our example networks, with the increase of the number of total edges m in the network, the density of the network increases, meaning that we do not observe a topology of connected nodes embedded atop the "background" of disconnected nodes any more. Instead, the number of parallel edges between an arbitrary pair of nodes increases, with the highest number being seven in the second network, versus three in the first network. This creates higher variability in the number of parallel edges and thus, a broader distribution of the value of Φ_{ij}.

Let us now identify whether there are pairs of nodes in our two networks that interact significantly more or less than predicted by the configuration model. To achieve this, we threshold the values of $\mathbf{\Phi}$ at level $\alpha = 0.05$. For the first network, we find three pairs

Instead, for each pair i, j there are Ξ_{ij} (cf. 3.2) discrete values of Φ_{ij}.

[2] The degree distribution is approximately binomial. Sum of independent binomials with the same success parameter is a binomial, however, in our case the degree is a sum of *dependent* binomials with the same success parameter $p = 1/(n^2 - n)$.

of nodes that exhibit over-represented interactions. For two of these pairs, the number of interactions equals three and is the maximum observed in the network. Note, however, that not all pairs with the maximum number of interactions are selected. For the second network, we find two pairs with over-represented interactions and three pairs with under-represented interactions. The latter are pairs that do not interact at all. As the networks are realisations of a $G(n, m)$ model, we would expect to observe only few pairs with under-represented or over-represented interactions which happen by mere chance. Recalling the relation between the threshold level α and a statistical significance level discussed in Section 4.1, we can say that $\alpha = 0.05$ approximately corresponds to 0.025 significance. This means that up to 2.5% of the node pairs, which is approximately 5 pairs, can be selected by our threshold by mere chance. In order to refine the procedure such that the error of detecting such significant pairs, one can apply the methods of *multiple hypothesis testing* [166, 178]. However, we leave this refinement to the future research.

Network with heterogeneous degree Let us now consider a random directed network realisation of a configuration model with $m = 500$ edges among $n = 20$ nodes with heterogeneous degrees that also allows self-loops. We draw the degrees from the log-normal distribution $\ln \mathcal{N}(\mu, \sigma)$ with $\mu = 2.5, \sigma = 0.9$. The network is shown in Fig. 4.5(a) by means of its adjacency matrix, with the nodes sorted according to the degree. As before, we apply the signed measure to all pair of nodes in this network based on the unbiased hypergeometric ensemble and we obtain the matrix shown in Fig. 4.5(b). Different from the previous example, there is a visible structure in the signed matrix. Namely, we see both positive and negative relations in the upper left corner, which corresponds to pairs of high-degree nodes that are almost all connected. From this corner that exhibits no particular pattern in values, there is a transition to almost uniformly neutral relations in the bottom right corner, which corresponds to the low-degree pairs. Along this diagonal, we observe that the relations between nodes that have edges in the original network become increasingly positive and that they are embedded in the "background" of negative relations between disconnected nodes. This negative background exhibits a smooth gradient from more negative values to more neutral values, which is due to the fact that zero edges between higher degree nodes map to a more negative value than zero edges between lower degree nodes. This phenomenon is also reflected in the histogram of values of $\mathbf{\Phi}$, shown in Fig. 4.5(c), where we see an approximately uniform distribution of positive values, a high peak near the neutral $\Phi = 0$ and a decreasing distribution of more negative values. At the

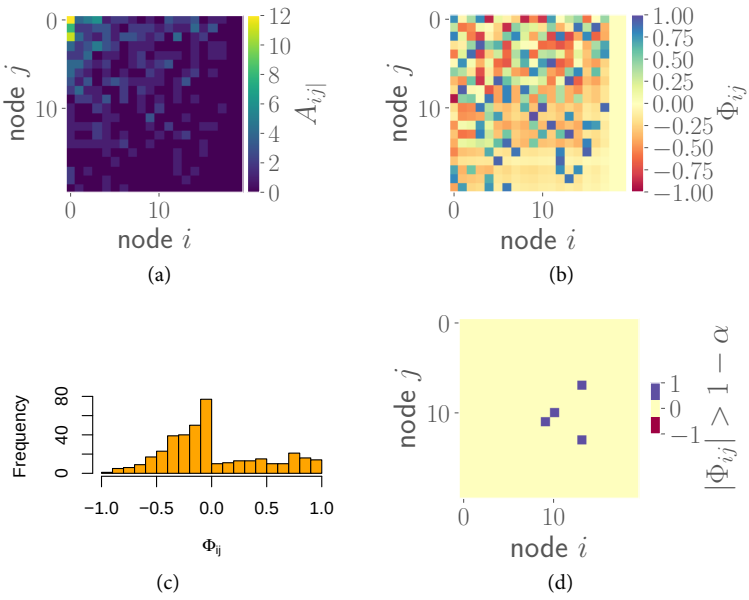


Figure 4.5: A network realisation of a configuration model with $m = 500$ multi-edges among $n = 20$ nodes with heterogeneous degree distribution. The adjacency matrix (a), (b) the matrix $\mathbf{\Phi}$ of the signed relations inferred based on the unbiased hypergeometric ensemble (cf. Eq. (3.2)), (c) the histograms of values of Φ_{ij} for all n^2 pairs of nodes, and (d) the node pairs with significantly under-represented ($\Phi_{ij} < -1 + \alpha$) and over-represented ($\Phi_{ij} > 1 - \alpha$) edges at threshold $\alpha = 0.05$.

same threshold level as before, $\alpha = 0.05$, we find only four pairs with over-represented edges, which corresponds to exactly 1% of all pairs in the network and which is below the number that we would expect to happen by chance at the given threshold level.

Network with block structure As the last synthetic example, we generate a random network with block (community) structure, with the nodes within a block having ten times higher probability to form an edge compared to nodes in different blocks. The $n = 20$ nodes are split into three blocks of an approximately equal size (six, six and eight) according to their degree. As in the previous example, the nodes have a heterogeneous degree distribution drawn from the log-normal distribution $\ln \mathcal{N}(\mu, \sigma)$ with $\mu = 2.5, \sigma = 0.9$, such that there are $m = 500$ total edges in the network. The adjacency matrix of one network realisation is shown in Fig. 4.6(a) and the corresponding signed matrix $\mathbf{\Phi}$ based on the unbiased hypergeometric ensemble in Fig. 4.6(b). Due to the

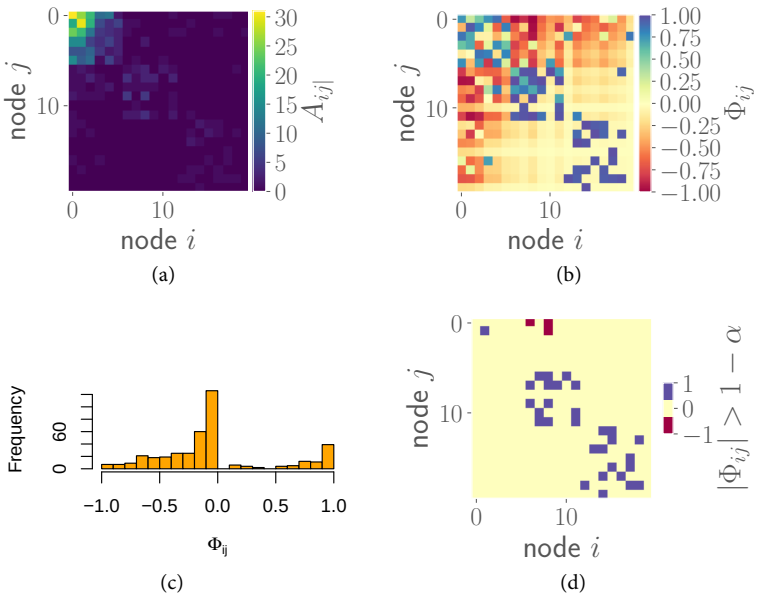


Figure 4.6: A network realisation of a block model with $m = 500$ multi-edges among $n = 20$ nodes with heterogeneous degree distribution and likelihood for an edge between nodes in the same block ten times higher than for an edge between nodes in different blocks. The adjacency matrix (a), (b) the matrix $\mathbf{\Phi}$ of the signed relations inferred based on the unbiased hypergeometric ensemble (cf. Eq. (3.2)), (c) the histograms of values of Φ_{ij} for all n^2 pairs of nodes, and (d) the node pairs with significantly under-represented ($\Phi_{ij} < -1 + \alpha$) and over-represented ($\Phi_{ij} > 1 - \alpha$) edges at threshold $\alpha = 0.05$.

higher concentration of edges in the block of highest degree nodes, the gradient of negative values similar to the previous example is more pronounced. Applying the threshold $\alpha = 0.05$ to this network results in three significantly negative relations and 30 significantly positive relations, which is 8.25% of all pairs in the network. This considerable percentage indicates that the unbiased hypergeometric ensemble (configuration model) is not a good fit to explain the network. However, it serves as a good null model to highlight the block structure in the network. Figure 4.6(d) shows that all highly positive values are within the blocks, with the two blocks of lower degree nodes being clearly identifiable. However, the block of the highest degree nodes is not seen in this filtered matrix because the signal of the block structure is concealed by the expectation from the configuration model, which already predicts high number of edges between the nodes in this block. From this example, we see that applying the signed relations provides yet another way to infer the backbone of the network. So far

we have only considered the unbiased hypergeometric ensemble as the model behind the deviations. However, we are not limited to that. We will show below on empirical examples how to use the general ensemble to account for diverse patterns in a network and to identify the deviations on top of these patterns.

4.2.2 Empirical networks

With our understanding of how our signed measure behaves based on the synthetic examples, we now apply it to two empirical networks.

Zachary's karate club First, we revisit the network of Zachary's karate club, which we studied in Chapter 3 [215]. To recall, it records 231 interactions between 34 members of the karate club, which are divided into two communities. Applying the measure Φ defined in Eq. (4.6) based on the unbiased hypergeometric ensemble leads to the matrix shown in the top left of Fig. 4.7, where the nodes are ordered according to their index in the original dataset. The index is ordered in such a manner that the first half of indices approximately corresponds to one community and the second half to the other. Also, nodes with a higher degree in the first community have a lower index (top left corner of the matrix) and nodes with higher degree in the second community have a larger index (bottom right corner of the matrix). Such ordering of nodes unveils the community structure in the signed matrix. We see two blocks of positive relations in the opposite corners of the matrix along the diagonal, i.e., within the communities, and blocks of negative relations in the two corners corresponding to relations across the communities. Filtering the original network based on the significant signed relations Φ_{ij} with a threshold $\alpha = 0.01$, i.e., applying the condition $|\Phi_{ij}| > 1 - \alpha$, leads to the network shown in the bottom left of Fig. 4.7. We obtain a backbone of the network with 20 connected pairs of nodes out of 78 pairs in the original network. What happens if we compute the signed relations based on a different null model instead of the configuration model? From Chapter 3 we know that Zachary's karate club is well described by the generalised hypergeometric ensemble with block-structured edge propensities. Applying the measure Φ to a network based on that ensemble results in the signed matrix shown in the top right of Fig. 4.7. Compared to the configuration null model, in this case we observe very few negative relations, which all are insignificant with their absolute values considerably smaller than one. The two pronounced blocks of strongly positive relations are also gone, being replaced by blocks with alternating positive and negative relations with comparatively low absolute

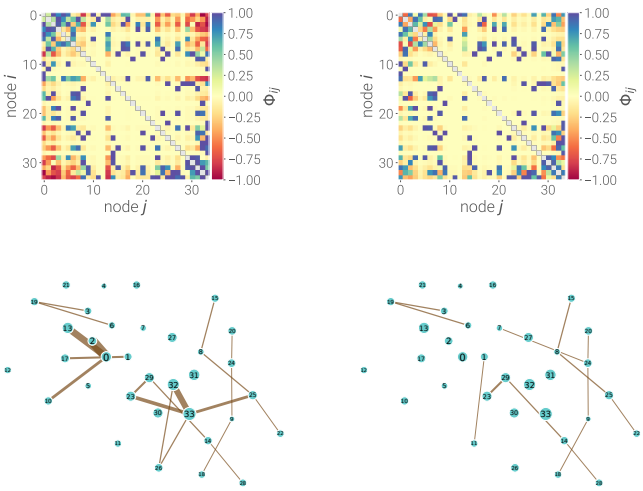


Figure 4.7: The network of Zachary's karate club (cf. Figs. 3.4 and 3.9). The signed matrices $\mathbf{\Phi}$ based on (top left) the unbiased hypergeometric ensemble and (top right) the hypergeometric ensemble with block propensities (cf. Eq. (3.6)). The filtered networks resulting from the threshold condition $\Phi_{ij} > 1 - \alpha$, with $\alpha = 0.01$ are shown in the bottom row for both matrices. There are no pairs with under-represented interactions at $\alpha = 0.01$ in both cases.

values. Applying a filter to the original network as before ($|\Phi_{ij}| > 0.99$), we obtain the network shown in the bottom right of Fig. 4.7. It has 12 connected pairs (compared to 8 obtained by the percentile-based filter in Section 3.3.2) out of 78 in the original network, all of which have low multiplicity of edges between them.

Eurovision song contest In our last example, we consider the network of votes between countries participating in the Eurovision[3] song contest between the years 1975 and 2015 [76]. The contest is an annual competition among countries that are member of European Broadcasting Union. Each participating country is represented by one song, which is performed at an event broadcasted live in at least each participating country. The countries then gather votes towards other countries (no self-voting) from the population by means of televoting and from a jury. These votes are then aggregated in each country, leading to a ranking of all other participating countries. This ranking is then turned into points for the top ten countries: 12 points to the top,

[3] `https://eurovision.tv/`, data retrieved and provided by David Garcia

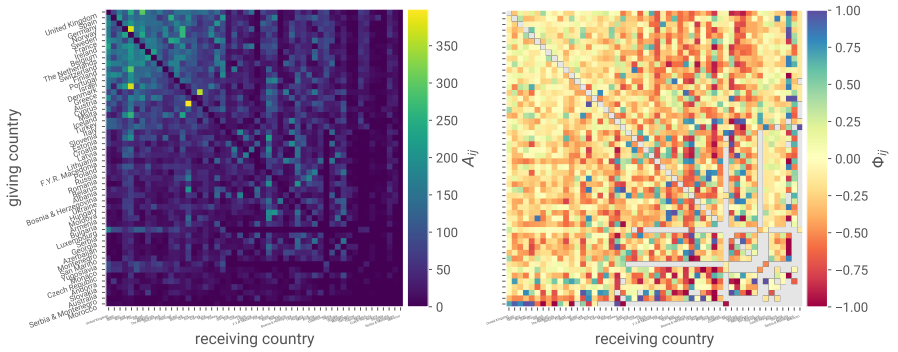


Figure 4.8: (Left) the adjacency matrix of the aggregate network of votes among countries participating in the Eurovision song contest between 1975–2015. The value A_{ij} is the sum of all points country i has given to the country j in this period. (Right) the corresponding signed relations Φ_{ij} based on the unbiased hypergeometric ensemble. The countries are ordered according to the total number of acquired points.

10 to the second ranked, and from 8 to 1 to the third to tenth ranked. Representing the votes as weighted edges between countries means that each country participating in a given year has ten outgoing edges with weights corresponding to the respective points.

We build an aggregate network of votes between the countries by summing all the points that one country gave another country in the 41 competitions between 1975 and 2015. The adjacency matrix $\hat{\mathbf{A}}$ is shown in the left panel of Fig. 4.8 and we apply to it the measure Φ based on the unbiased hypergeometric ensemble. Note that by doing so, we approximate the voting procedure by assuming that each point can be distributed freely by a given country. For instance, our null model allows, in principle, for a country in a given year to allocate all the 58 points to one country, without giving any points to others. We argue, however, that this approximation is valid for the aggregate of 41 years, in which there is a wide distribution of possible points each country can give to another.

Let us now compare the signed relations that we obtained with the ones by García and Tanase [76]. The authors of [76] compute the "Friend-or-Foe" (FoF) coefficient between countries i and j for the year t according to

$$FoF_{ij}(t) = \frac{\hat{A}_{ij}(t)}{12} - \frac{\sum_{l\neq i}\hat{A}_{lj}(t)}{12(n-2)}, \tag{4.10}$$

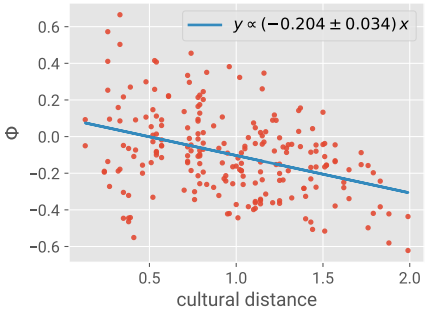


Figure 4.9: The signed relations and the Hofstede's cultural distance between countries that co-participated in the Eurovision song contest at least 25 times. The blue line shows the result of a linear regression with $R^2 = 0.146$ and p-value $= 5.8e-9$.

which is the normalised difference of the points from i to j and the average points that the country j received from all other countries. To construct the aggregate signed network between the years 1975 and 2012, the authors take the average FoF coefficient between each pair of countries, $\langle FoF_{ij} \rangle = \sum_t FoF_{ij}(t)$ for each year t that i and j co-participated in the contest. The authors then compare the resulting mean FoF coefficient to the *cultural distance* between the countries. To define the cultural distance, they use the quantification of cultural distances provided by Hofstede [91], which includes four cultural dimensions: Power Distance *cp*, Individualism *ci*, Masculinity *cm*, and Uncertainty Avoidance *cu*. The four dimensions are aggregated into one dimensional distance by means of Manhattan distance:

$$d_{ij} = \frac{1}{100}(|cp_i - cp_j| + |ci_i - ci_j| + |cm_i - cm_j| + |cu_i - cu_j|). \tag{4.11}$$

To have a reliable mean FoF, the authors consider only the countries that co-participated in the contest at least 25 times. Replicating this procedure for our signed network $\mathbf{\Phi}$ results in 216 pairs of countries, for which we compare Φ_{ij} to the cultural distance d_{ij}, as shown in Fig. 4.9. We find a negative relation between the two variables, with Pearson correlation of -0.383 and the linear regression slope of 0.204 ± 0.034 ($R^2 = 0.146$ and p-value $< 10^{-8}$). So, the further two countries are in terms of culture, the lower the signed relation between them. The relation tends towards negative values starting from cultural distance $d_{ij} \approx 0.5$. Our result is in line with the result of García and Tanase, who find a Pearson correlation of -0.441 and the linear regression slope -0.0956 (computed on binned data). One major advantage of our method is that it is an unbiased statistical measure, so the results must be reliable also for the countries

Table 4.1: The most under-voting and over-voting pairs of countries in Eurovision song contest aggregated between 1975 and 2015. Only the pairs co-participating at least 25 times in the contest are considered.

Most under-represented		Φ
Cyprus	Turkey	-0.635
Turkey	Cyprus	-0.633
Denmark	Greece	-0.622
Greece	Sweden	-0.581
Portugal	Turkey	-0.551
Denmark	Spain	-0.516
Ireland	Turkey	-0.507
Austria	Cyprus	-0.499
Turkey	Denmark	-0.498
Ireland	Greece	-0.467
France	Malta	-0.465
Sweden	Spain	-0.449
Greece	Turkey	-0.446
Malta	France	-0.445
Greece	Germany	-0.443
Turkey	France	-0.442
Greece	Denmark	-0.436
Greece	Israel	-0.422
Sweden	Greece	-0.422
Spain	Switzerland	-0.420

Most over-represented		Φ
Cyprus	Greece	0.978
Greece	Cyprus	0.797
Denmark	Sweden	0.665
Norway	Sweden	0.573
Norway	Iceland	0.505
Sweden	Denmark	0.503
Sweden	Iceland	0.482
France	Portugal	0.455
Iceland	Sweden	0.431
Malta	United Kingdom	0.428
United Kingdom	Ireland	0.426
Norway	Denmark	0.415
Sweden	Norway	0.411
Spain	Portugal	0.407
The Netherlands	Belgium	0.393
Switzerland	Spain	0.382
Iceland	Norway	0.362
Spain	Germany	0.351
Denmark	Germany	0.346
Finland	Israel	0.334

that did not co-participate many times. In Fig. A.1, we recompute the relation between Φ_{ij} and d_{ij} between all countries and we obtain a weaker but still significant result of Pearson correlation -0.214 and linear regression slope of 0.150 ± 0.024 ($R^2 = 0.046$ and p-value $< 10^{-8}$).

Next, in Table 4.1 we show the lists of countries with the lowest 20 and the highest 20 values of Φ_{ij} among the ones that co-participated in the Eurovision song contest at least 25 times. Remarkably, both lists are topped by reciprocal relations involving Cyprus. It has the most positive relation with Greece and the most negative relation with Turkey. The former reflects the fact that the Republic of Cyprus is inhabited mostly by ethnic Greeks and the latter is a clear mark of the "Cyprus dispute"—the unresolved issue of the Turkish military occupation of a part of the island since 1974. In Table A.1, we present a similar table that accounts for all countries.

Last but not least, let us consider the reciprocity of the relations between countries. In order to quantify the overall reciprocity in an unsigned unweighted network, Gar-

laschelli and Loffredo proposed to use the correlation coefficient between the symmetric entries of the adjacency matrix [77]:

$$\rho = \frac{\sum_{i \neq j} (\hat{A}_{ij} - \langle \hat{\mathbf{A}} \rangle)(\hat{A}_{ji} - \langle \hat{\mathbf{A}} \rangle)}{\sum_{i \neq j} (\hat{A}_{ij} - \langle \hat{\mathbf{A}} \rangle)^2}, \tag{4.12}$$

where $\langle \hat{\mathbf{A}} \rangle = m/(n^2 - n)$ is the density of the edges in the network. This measure of reciprocity takes values in the range $\rho \in [-1, 1]$, with $\rho = -1$ denoting the extreme case where all pairs anti-reciprocate—if there is an edge (i, j) then there is no edge (j, i)—and $\rho = 1$ denoting maximum reciprocity, i.e., for any edge (i, j) there is also an edge (j, i).

Technically, Eq. (4.12) can be applied also to weighted or multi-edge networks. Then, the difference $\hat{A}_{ij} - \langle \hat{\mathbf{A}} \rangle$ measures the deviation of the observed number of interactions from the average in the network. Hence, we can straightforwardly substitute that difference by our signed relations Φ_{ij}, which measure the deviation of the observed interactions from the prediction of the null model in a more sophisticated manner. We arrive at the following definition of reciprociy in the network:

$$\rho = \frac{\sum_{i \neq j} \Phi_{ij} \cdot \Phi_{ji}}{\sum_{i \neq j} \Phi_{ij}^2}, \tag{4.13}$$

We can also adapt it such that only the reciprocity in the signs and not the absolute values of the relations Φ_{ij} are captured as follows:

$$\rho^{\text{sign}} = \frac{\sum_{i \neq j} \text{sign}\left(\Phi_{ij} \cdot \Phi_{ji}\right)}{n(n-1)}, \tag{4.14}$$

with $\text{sign}(x) = 1$ if $x > 0$, $\text{sign}(x) = -1$ if $x < 0$ and $\text{sign}(x) = 0$ if $x = 0$. So, if the values of the relations Φ_{ij} and Φ_{ji} are both greater or both less than zero, the pair i and j contributes positively to the reciprocity of the network. Otherwise, they contribute negatively. As before, the values of reciprocity defined in Eq. (4.14) fall in the range $\rho^{\text{sign}} \in [-1, 1]$.

We now apply the two definitions of reciprocity to the signed relations between countries resulting in $\rho = 0.524$ and $\rho^{\text{sign}} = 0.681$. Both numbers show that the average voting behaviour in the contest tends to be reciprocal. We further illustrate this in

Table 4.2: Counts of node pairs for three types of reciprocation measured at different threshold α, i.e., a pair is counted if both $|\Phi_{ij}| > \alpha$ and $|\Phi_{ji}| > \alpha$. The threshold in the middle column is equal to the mean of the the signed matrix, $\langle \mathbf{\Phi} \rangle = -0.153$.

α:	0	$\lvert\langle \mathbf{\Phi} \rangle\rvert$	0.5
$++$	183	100	40
$-+$	343	115	10
$--$	691	440	66

Table 4.2, where we count the reciprocating ($++$ and $--$) and anti-reciprocating ($-+$) pairs of nodes. We account only for the pairs for which the absolute values of the relations in both directions are above a certain threshold, $|\Phi_{ij}| > \alpha$ and $|\Phi_{ji}| > \alpha$. That means, the larger the α, the stronger both relations must be for the pair to be counted as reciprocating or anti-reciprocating. We find that the fraction of anti-reciprocating pairs in the Eurovision song contest decreases with growing α: from 28.2% for $\alpha = 0$, to 17.6% for $\alpha = |\langle \mathbf{\Phi} \rangle| = 0.153$, to only 8.6% for $\alpha = 0.5$. In other words, if two countries both have a strong bias towards each other, then this bias tends to be in the same direction, positive or negative, for both. Most of the ten strongest anti-reciprocating pairs include a country with very few participations in the contest, except the pair Armenia–Turkey ($\Phi_{A,T} = -0.617, \Phi_{T,A} = 0.910$), which is in agreement with the results of García and Tanase [76].

4.3 Conclusion

We have developed a tool that measures the deviation of a given value from the median of a distribution. In contrast to existing measures, ours is non-parametric and is valid for a wide range of distributions, which can be skewed, bounded, continuous or discrete. This allowed us to infer signed relations (network) from unsigned counts of repeated interactions.

Because the measure is unbiased, we were able to compare the signed relations between pairs of nodes with strongly varying degrees. This is different from the previous attempts to infer signed relations from unsigned networks, where only nodes with a certain minimum degree could be compared [76].

As mentioned earlier, there is interest in signed social relations (e.g., friendship versus animosity), but there is a lack of data on signed social networks. We believe that our

methodology can mitigate this problem by constructing new data sets on signed social networks from the abundant data on unsigned networks.

We showed how filtering only highly significant signed relations provides a new way to infer a network backbone. While we understand how the threshold used on the values of signed relations maps to statistical significance for a continuous variable, we have left it to the future research to understand this connection for discrete variables.

Chapter 5

Friend or Foe? Significant deviations in citation behaviour

Summary

We study the citation networks between most cited authors in multiple empirical data sets. Using the methodology from previous chapters, we first evaluate different formulations of topical similarity between authors in terms of their power to explain the empirically observed citations. Based on a generalised hypergeometric ensemble that encodes the best fitting similarity formulation and a temporal order preserving combinatorial component, we infer the extent of over-citations and under-citations among the most cited scientists. We find that the resulting signed relations are on average reciprocal between pairs of authors, with the patent networks exhibiting stronger reciprocity than networks in science. We also outline an approach to identify interdisciplinary pioneers and citation cartels.

This chapter has been written specifically for this dissertation. Based on unpublished discussions by Nanumyan V., Casiraghi G., Mavrodiev P., Schweighofer S., Scholtes I., and Schweitzer F. VN performed the analysis and interpreted the results.

In Chapter 3, we developed a novel statistical ensemble of networks that allows modelling networks with heterogeneous preferences of nodes for choosing with whom to form edges. In Chapter 4, we formulated an unbiased statistical measure that allows quantifying how much and in which direction an observed value deviates from a given distribution. Using the new measure and the ensemble, we showed how to infer a signed network from observations of unsigned repeated interactions between nodes. In this chapter, we apply the methodology of signed network inference to collaborative knowledge networks, in order to quantify how the citations between authors deviate from the prediction based on purely scientific criteria.

In Section 1.2 we briefly discussed three social biases—homophily, reciprocity and structural balance—that universally affect human social relations including citations and collaborations among scientists. There are, however, other social biases, which are specific to the academic community, namely the *selection bias* in recognising the scientific work of peers, and the *evaluation bias* in reviewing the scientific work of peers [171]. These biases are aggravated by competition for scarce resources within the scientific community, namely *attention* and *funding*. In addition, the exponential growth of the number of publications leads to the problem of information overload. Scientists, unable to keep track of the developments in their research areas, become selective in their recognition of peers. Even within their topical communities, they tend to cite works that are *already* highly cited, works of scientists they already know (e.g., through collaborations or conferences), or works of scientists that have previously cited them (reciprocity). All this reinforces existing structures and creates *filter bubbles* that limit the diversity of information perceived by individuals. Furthermore, competitive funding creates incentives for *strategic citation behaviour* of scientists, either *by not recognizing* the scientific work of peers that compete for the same funds or *by negatively evaluating* their manuscripts or proposals. While such behaviour is unethical, it is very hard to detect and to avoid in a system that inherently builds on a *peer review process*.

A number of works have highlighted the importance of social structures and processes, in particular: (i) the effect of collaborations and *social acquaintance* on citation structures [165], (ii) the influence of a *small elite of scientists* on the knowledge generation process [7], (iii) the *first-mover advantages* in the evolution of citation networks [135], (iv) the effect of *attention mechanisms* towards scientific publications [149], and (v) the role of *elite institutions* in funded research [116].

The discipline of scientometrics has dealt with collaborative knowledge networks since

at least the middle of the twentieth century [159]. These studies mostly focused on identifying important individuals by means of established topological measures, such as centralities [71, 72, 136, 138, 145]. Only in a very recent study, missing citations have been identified between closely related physics publications [42]. However, similarity between articles was only calculated based on the overlap between their bibliographies, which does not provide a sound null model.

Our methodology of inferring signed relations from unsigned data enables us to go beyond such investigations. Precisely, it will allow us to quantify interpersonal relations that are addressed in the social sciences [66], but could hardly be measured. Detecting and measuring social biases will also permit us to shed new light on the *origin* of such biases [54]. To achieve this goal, we first refine the ensemble formulation to better represent the formation mechanisms of collaborative knowledge networks. In the following, we define a new *temporal order preserving configuration model* for citations between scientists, which accounts for the temporal sequence of publications. We show how this new model can be used to infer the backbone of citations between the most prominent authors in a given scientific community. Then, we discuss how to encode the topical similarities between scientists in the null model. This is needed because scientists are more likely to cite the works that are topically closer to their work, due to purely scientific criteria. We compare four different formulations of the topical similarity by means of their statistical power to explain the observed citations between the most cited authors. We use the best model to infer the signed relations between these authors. Finally, we analyse how these signed relations correlate with such characteristics of the authors as the total number of their citations, or the number of previous co-authorship relations between authors.

5.1 Modelling citations among top authors

Collaborative knowledge networks grow over time. In particular, the citation network among knowledge artifacts $G^{pc} = (V^p, E^{pc})$ (cf. Section 2.2) grows by adding new artifacts which cite other artifacts that are already in the network. As a result, the network G^{pc} is a *directed acyclic graph* (we disregard rare exceptions when an older knowledge artifact cites a newer one). If we model this network by means of network ensembles that we have discussed so far, the property of having no cycles—the temporal constraint—will not be preserved. Both the standard Molloy-Reed configuration model (Section 1.2) and our hypergeometric ensembles (cf. Eqs. (3.1) and (3.5)) built

on the same assumptions as the configuration model, do not respect the temporal order in which the nodes enter the network. For a citation network, this means that the models allow for edges from older nodes to newer nodes. In principle, it is not difficult to limit the models and their probability spaces post hoc, by removing the model realisations that break the temporal constraint. However, this approach is computationally costly and it removes the benefits of the analytical tractability of the hypergeometric ensembles. The problem is exacerbated if we consider the citation network G^{ac} between authors (the projection of G^{pc} into a layer $G^{ac} = (V^a, E^{ac})$, cf. Section 2.2). The edges of the network G^{ac} are still temporally constrained, as an author who published earlier cannot cite an author who published later. In this case, however, the network is not acyclic and post hoc adjustments of the models are not feasible. In the following, we show how the formulation of a generalised hypergeometric ensemble can be broadened to respect the temporal constraints of growing networks and their projections.

5.1.1 Temporal order preserving configuration

The generalised hypergeometric ensembles developed in Chapter 3 build on two components: the combinatorial one represented by the matrix $\boldsymbol{\Xi}$ and the one that encodes the heterogeneous preferences among pairs of nodes to form an edge, represented by the matrix of edge propensities $\boldsymbol{\Omega}$. So far, we have only encoded the configuration model as the combinatorial component of the ensemble by setting $\Xi_{ij} = k_i^{\text{out}} k_j^{\text{in}}$, where $k_i^{\text{out/in}}$ is the out-degree or in-degree of node i. However, thanks to the definition of the ensemble, we are not limited to the configuration model.

To show how we can redefine the matrix Ξ to incorporate temporal constraints in the ensemble, let us consider a small example of a collaborative knowledge network. The left panel of Fig. 5.1 shows a citation network between ten knowledge artifacts written by five authors. The nodes are plotted on a timeline according to the time at which they are added to the network. The citations drawn from newer to older nodes all satisfy the causal constraint imposed by the temporal order of nodes. The projection of this network into author–author citations is shown in the right panel of Fig. 5.1. This projection has loops, e.g., between nodes $A0$ and $A1$, as well as a self-loop belonging to node $A1$.

It is not clear from observing just the author-author citation network in Fig. 5.1 that there cannot possibly be citation edges from authors $A0$, $A1$, $A2$ to authors $A2$, $A3$ and

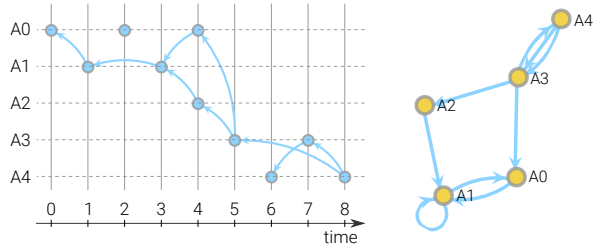


Figure 5.1: (Left) time-unfolded network of citations between publications written by five authors and (right) the corresponding projection to author–author citations.

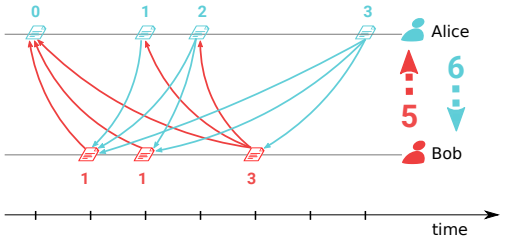


Figure 5.2: All possible citation edges between time-stamped publications of two authors.

$A4$, given the temporal sequence of their publications. If we define an ensemble based on configuration model, it would allow, e.g., up to $\Xi_{A1,A3} = 2 \cdot 2 = 4$ citations from $A1$ to $A3$. How can we fix this?

Figure 5.2 shows the publication timelines of two authors, Alice and Bob, and all possible citations that one's publications can make to the other's. That is, each publication of one author can cite all the preceding publications of the other. The total possible number of those is shown in the figure next to the publication. Given the publication timelines, Alice can cite Bob six times, while Bob can cite Alice 5 times. We can use these numbers to define the *temporal order preserving possibility matrix* $\boldsymbol{\Xi}$ for the hypergeometric ensemble of citations between authors, which formally writes as

$$\Xi_{ij} = \sum_t \delta\big(T(v), t\big)\, \delta\big(B^a_{iv}(t), 1\big) \sum_l B^a_{jl}(t), \quad \text{for } v \in V^p, \tag{5.1}$$

where $\mathbf{B}^a(t)$ is the incidence matrix of the authorship relations at time t (cf. Section 2.1.1), $T(v)$ is the time at which publication v was added to the network, and the Kronecker delta $\delta(x, y) = 1$ if $x = y$ and is zero otherwise.

Let us now compare the hypergeometric ensemble, Eq. (3.1), based on Eq. (5.1), to the

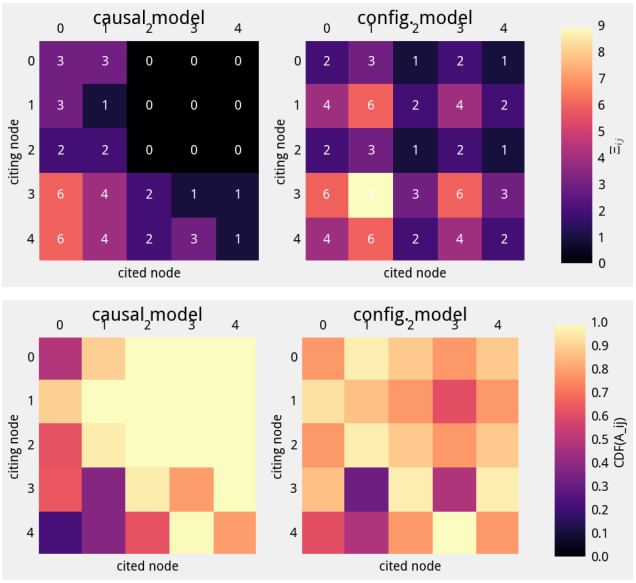


Figure 5.3: (Top) the matrix of possible edges according to Eq. (5.1) and according to configuration model, and (bottom) the corresponding marginal probabilities of the observed edges according to Eq. (3.2) for the author–author citation network shown in Fig. 5.1.

one based on the configuration model, $\Xi_{ij} = k_i^{\text{out}} k_j^{\text{in}}$. Figure 5.3 shows in the top panel the two matrices Ξ for our toy example. As expected, the order preserving matrix does not allow any edges from nodes $A0, A1, A2$ to nodes $A2, A3$ and $A4$. The effect of the temporal constraint on the network ensemble can be seen in the bottom panel of Fig. 5.3, where we show the marginal probability (Eq. (3.2)) of the observed citations in our toy example.

5.1.2 Author similarity

Authors are not expected to cite other authors uniformly. As discussed in Chapter 1, it is expected that authors who are more similar in their research topics cite each other more frequently. Our generalised hypergeometric ensemble allows encoding of these similarities by means of the matrix $\mathbf{\Omega}$ of edge propensities, as defined in Eq. (3.5).

The topical similarities between authors can be defined based on the similarities of their publications. This can be done, for instance, by means of keyword similarity, where the keywords are either provided in the metadata of the publications, or ob-

tained by applying natural language processing tools to the content of publications [3]. Another approach to quantifying the similarity of publications is based on the properties of the collaborative knowledge network of which they are a part [8, 32, 57, 58, 59, 212].

Below we describe two commonly used similarity measures for publications, *bibliographic coupling* and *co-citation similarity*. Although, they are mostly used to measure publication similarity, they are also applied to collaborative knowledge networks to directly measure the similarity between authors, as we will show below.

Bibliographic coupling was introduced in 1963 by Kessler [101]. Two given publications are bibliographically coupled if they both cite the same publication. The coupling strength is defined by the number of such publications. It can be defined for authors as well, with two authors being bibliographically coupled if their publications cite the same publication [216]. For publications, bibliographic coupling is a retrospective and static measure, meaning that it is based on information available at the time of publication of the two compared articles and it does not change over time. However, bibliographic coupling is not static for authors, as it changes when they write new publications.

Co-citation similarity was introduced independently by Small and Marshakova in 1973 [119, 185]. For two publications, it measures how often they are both cited in the same publication. This measure addresses the above-mentioned drawback of the static and retrospective bibliographic coupling between publications, as it evolves over time with newly published articles, possibly increasing the co-citation between the two publications. Co-citation similarity is also defined for authors based on their publications [207].

The two similarity measures are considered as proxies of semantic, or topical, relatedness based on the assumption that the more two publications cite or are cited by the same publication, the more similar is their context or content. From the network perspective, one can see bibliographic coupling of two nodes as an outward measure, as it is based on the outgoing edges of the two nodes, and co-citation similarity as an inward measure, as it is based on the incoming edges of the two nodes.

While the two similarity measures can be defined by means of absolute counts, it is useful to normalise them with respect to the total citations that the two considered

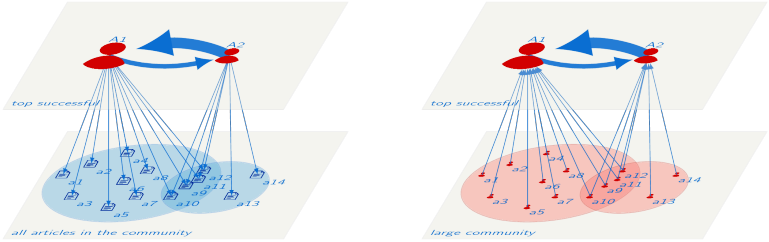


Figure 5.4: (Left) bibliographic coupling between two authors based on the publications they write and (right) co-citation similarity of the two authors based on citing authors.

publications or authors make or receive. For this, *cosine similarity* [61] or *Jaccard index* are commonly used [60]. Cosine similarity is defined as the cosine of the angle between two vectors. Jaccard index is defined for two mathematical sets as the fraction between the intersection and the union of the sets and is defined for two nodes in a simple graph in Eq. (2.7). The bibliographic coupling (outward) and the co-citation similarity (inward) for two publications write in terms of the Jaccard index as

$$\sigma_{ij}^{out} = \frac{\sum_l I[A_{il}A_{jl}]}{k_i^{\text{out}} + k_j^{\text{out}} - \sum_l I[A_{il}A_{jl}]}, \tag{5.2}$$

$$\sigma_{ij}^{in} = \frac{\sum_l I[A_{il}A_{jl}]}{k_i^{\text{in}} + k_j^{\text{in}} - \sum_l I[A_{il}A_{jl}]}, \tag{5.3}$$

$$\tag{5.4}$$

where A_{ij} are the entries of the adjacency matrix of the citation network G^{pc} among publications, the function I equals one for a non-zero argument (similar to Heaviside step function, except for $I[0] = 0$). Similarly, we can calculate the similarities between authors by substituting the adjacency matrix of G^{pc} by the adjacency matrix of the appropriate projections of the collaborative network (author–publication citation network for bibliographic coupling and publication–author citation network for co-citation similarity) and with the degrees that disregard the multiplicity of edges, i.e., $k_i^{\text{out}} = \sum_j I[A_{ij}]$. Moreover, we can compute the similarities based only on author–author relations, if we use the projection G^{ac} of citations onto authors. In the example shown in the left of Fig. 5.4, the bibliographic coupling between two authors based on the publications they cite is equal to 2/7 as they both cite four publications out of 14 total. The right panel of Fig. 5.4, shows an example of co-citation similarity of two authors based on citing authors.

5.1.3 Edge propensity from author similarity

With the topical similarities between authors defined, we now use these in our ensemble for citations among authors by setting the edge propensities based on the topical similarities, $\Omega_{ij} \sim \sigma_{ij}$. By definition, the similarity based on Jaccard index is symmetric, $\sigma_{ij} = \sigma_{ji}$. However, the network ensemble for citations is directed, meaning that the edge propensities must be directed too. Hence, we need to redefine the similarity as an asymmetric measure to account for the directionality in the network.

Asymmetric similarity Let us try to understand how likely two authors $A1$ and $A2$ are to cite each other. Assume $A1$ is an established and prolific author, writing many publications in a broad spectrum of topics, and $A2$ is a young researcher focused on one particular topic. If the topic of $A2$ is among the topics of $A1$, then it is reasonable to assume that $A2$ should cite $A1$ more often than vice versa. We can quantify this by means of the fraction of citations that each of the two authors makes to the other, relative to all citations she makes, which writes as

$$\sigma_{ij}^{out} = \frac{\sum_l I[A_{il}A_{jl}]}{k_i^{\text{out}}}, \tag{5.5}$$

where A_{il} are the entries of the adjacency matrix of the author–publication projection of the citation network. $k_i^{\text{out}} = \sum_l I[A_{il}]$ is the corresponding degree that disregards the multiplicity of edges, i.e., it is the number of unique publications that author i cites. The asymmetric similarity measure in Eq. (5.5) is also called *Partial Jaccard coefficient* for arbitrary mathematical sets [75]. Equation (5.5) can be seen as an asymmetric modification of the bibliographic coupling between authors. For the example shown in the left of Fig. 5.4, $\sigma_{A1,A2}^{out} = 1/3$ and $\sigma_{A2,A1}^{out} = 2/3$. If we use these similarities as edge propensities in the ensemble, this would imply that $A2$ is twice as much likely to cite $A1$ than vice versa (everything else equal).

Similarly, we could estimate the likelihood of the two authors to cite each other based on the citation behaviour of the scientific community towards them. If, for instance, an arbitrary author cites author $A1$ every time when she cites $A2$ (but not vice versa), we could conclude that the works of $A2$ are closely related to, or even dependent on the works of $A1$. Hence, we would also expect $A2$ to cite $A1$ frequently (but not necessarily

vice versa).

$$\sigma_{ij}^{in} = \frac{\sum_l I[A_{li}A_{lj}]}{k_i^{\text{in}}}, \tag{5.6}$$

where A_{il} are the entries of the adjacency matrix of the author–author citation network and $k_i^{\text{in}} = \sum_l I[A_{li}]$ is the number of authors that cite the author i. Equation (5.6) can be seen as an asymmetric modification of the co-citation similarity between authors. Similar to the symmetric definitions of the two similarity measures in Eqs. (5.3) and (5.4), the asymmetric formulations can also be defined based on both publications and authors. In the following, we will refer to the similarity defined in Eq. (5.5) as `pub-out` if it is computed based on the cited publications and as `aut-out` if it is computed based on the cited authors. We will refer the to the similarity defined in Eq. (5.6) as `pub-in` if it is computed based on the cited publications and as `aut-in` if it is computed based on the cited authors.

Fitting propensities All the components are ready for defining a network ensemble that will serve as a null model for citations among authors. We defined in Eq. (5.1) the combinatorial component Ξ that preserves the temporal constraint on the possible citations imposed by the temporal order of publications. In Eqs. (5.5) and (5.6), we defined asymmetric topical similarities $\sigma_{ij}^{\text{in/out}}$ of author i to author j, which we can use as the edge propensity Ω_{ij}, such that the closer the authors are, the more likely they are to cite each other in the realisations of the ensemble. However, there are two problems with simply setting $\Omega_{ij} = \sigma_{ij}^{\text{in/out}}$. Firstly, the similarity $\sigma_{ij}^{\text{in/out}}$ can be equal to zero (and in practice it often is). Thus, the corresponding edge propensity would also be zero, meaning that the ensemble would not allow any edges to be drawn from i to j. This puts an implausible limitation on the ensemble, as in reality two authors can cite each other, even if they are not co-cited or bibliographically coupled. Secondly, it is not clear which similarity measure is the best fit for modelling the citations between authors, and setting propensity equal to a similarity does not allow for robust comparison and selection between different similarity definitions.

To overcome both problems, we use a technique for inferring propensities from (multiple) relation types by means of *multiplex network regression* based on the generalised hypergeometric network ensemble [35]. The method works as follows. Given r different types of positive-valued relations between the n nodes (an r-layer multiplex), which we suspect can have an effect on the formation of edges in the network (layer)

that is being modelled, we write the edge propensities as

$$\Omega_{ij} = \prod_{l=1}^{r} R_{l,ij}^{\beta_l^{(\Omega)}}, \tag{5.7}$$

where $\mathbf{R}_l$ is the weighted adjacency matrix of the l-th layer. That is, we set the propensity matrix entrywise product—also known as the Hadamard product—of the r adjacency matrices with exponentiated entries. For the adjacency matrix $\hat{\mathbf{A}}$ of the network being modelled, the ensemble defined by Eqs. (3.5) and (5.7) is a function f,

$$\Pr(\hat{\mathbf{A}}) = f(\mathbf{R}_1, \dots, \mathbf{R}_r; \beta_r^{(\Omega)}, \dots, \beta_r^{(\Omega)}), \tag{5.8}$$

relating the dependent variable, the probability of the network $\hat{\mathbf{A}}$, to the explanatory variables, $\mathbf{R}_l$. The exponents $\beta_l^{(\Omega)}$ are then fitted by maximum likelihood estimation (Eq. (7) in [35]), and the significance of the layer l is identified by a likelihood ratio test between the model with and without the layer (they are nested, as the model without the layer is achieved by setting $\beta_l^{(\Omega)} = 0$).

We apply the multiplex network regression to fit the edge propensities for citations between authors based on topical similarity of authors as follows. We build two matrices $\mathbf{R}_0$ and $\mathbf{R}_1$ from the matrix $\boldsymbol{\sigma}$ corresponding to one of the similarity definitions `pub-out`, `aut-out`, `pub-in` or `aut-in` as

$$R_{0,ij} = \begin{cases} 1 & \text{for } \sigma_{ij} \neq 0, \\ \varepsilon & \text{otherwise, with } \varepsilon \in (0,1), \end{cases} \tag{5.9}$$

$$R_{1,ij} = \begin{cases} \sigma_{ij} & \text{for } \sigma_{ij} \neq 0, \\ 1 & \text{otherwise.} \end{cases} \tag{5.10}$$

The matrix $\mathbf{R}_0$ is assigned a non-zero value ε to all its entries that correspond to the zero entries in the similarity matrix and ones to the elements where the similarity is non-zero. Similarly, the values in $\mathbf{R}_1$ are equal to values of the similarity where the latter are non-zero, and are one where the similarity is zero. Plugging the two matrices into Eq. (5.7) results in a formulation for the propensities in which a constant value $\varepsilon^{\beta_0^{(\Omega)}}$ is fitted instead of the zero entries of the similarity. The non-zero elements of the similarity are scaled to $\sigma_{ij}^{\beta_1^{(\Omega)}}$. By fitting the parameters $\beta_0^{(\Omega)}$ and $\beta_1^{(\Omega)}$, (i) we obtain the best representation of the citation propensities between authors based on their topical similarities, and (ii) we learn about the significance of the similarities in explaining

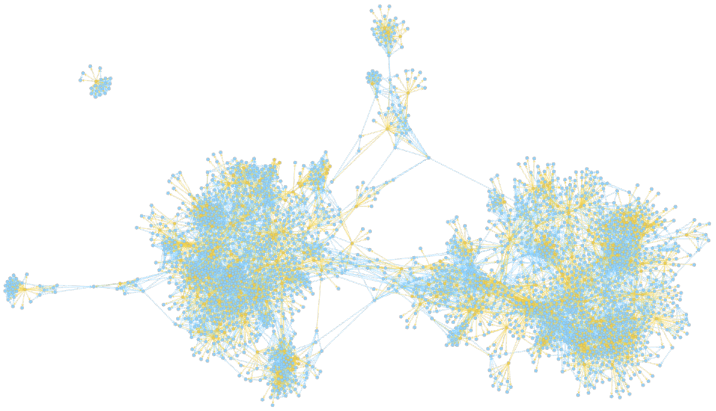


Figure 5.5: Citations among publications (blue) by the 200 most cited authors (yellow) and the authorship relations between authors and publications in the collaborative knowledge network of High Energy Physics in Physical Review Journals (PR-HEP).

the network of citations between the authors. This also means that we can compare different definitions of similarities in terms of their explanatory power for the observed citations. We cannot use the likelihood ratio test, as the models corresponding to different similarity definitions are not nested, but we could compare the Mahalanobis distances of the corresponding ensembles (cf. Section 3.2). Alternatively, we can use the computationally cheaper Akaike Information Criterion (AIC), because the parameters of the network regression are fitted by the maximum likelihood estimation.

5.2 Empirical networks of author–author citations

In the following, we apply the methodology described above to the twelve empirical collaborative knowledge networks introduced in Section 2.3. Specifically, we build the networks of the knowledge artifacts published within a time window of ten years: from January 1, 1960 to January 1, 1970 for the PR (which was discontinued in 1970) and from January 1, 2000 to January 1, 2010 for the others. Limiting the range of publication dates means that we compare authors and inventors that were writing publications and patents over the same time period, thus avoiding to give advantage to old researchers with many publications. It also reduces the effects of long-term trends in the evolution of the network on the outcomes of the analysis. Such time range limitation is similar to taking one temporal layer in the *dynamic configuration*

model proposed very recently [162]. The rate of new publications grows over time for all studied network, so taking the latest time window in the data sets results in the largest network for the given window length. For instance, the network of PR-HEP contains 10664 publications in the chosen time window out of 44829 in the full data set (cf. Table 2.2), or 23.8% of all publications. We analyse only 200 most cited authors in each network. However, we use the whole network when computing author similarities or aggregate properties, such as an author's total number of citations. Figure 5.5 shows the network visualisation of the citations between articles published in the ten year period by these 200 authors in PR-HEP, as well as the authorship relations between authors and publications. The figure highlights the existence of two topical communities in the network. While we are not limited technically to such small networks, we believe that focusing on the relations between the most prominent authors is especially important. One reason for this is the competition for limited resources, as mentioned in the beginning of this chapter. We believe that the competition is stronger among more prominent the researchers. For example, when competing for funding, these researchers are the ones who apply for the largest and most competitive funding. The peers who evaluate their applications are also selected from the small pool of prominent authors—who may also indirectly compete for the same funding—in order to match the level of expertise of the applicant.

Selecting the similarity measure Above we defined four network-based ways to measure the topical similarity between authors. Let us now fit and compare these measures for the twelve empirical networks. Table 5.1 shows the outcomes of performing multiplex network regression for the journals in INSPIRE data set, based on Eqs. (5.7) and (5.9) for the similarity definitions `pub-out`, `aut-out`, `pub-in` and `aut-in`. Tables B.1 and B.2 on Tables B.1 and B.2 show the same information for the patent and APS data sets. The AIC for each model is shown along with the two exponents $\beta_0^{(\Omega)}$ and $\beta_1^{(\Omega)}$ for each similarity definition. The similarity best describing the observed citations among authors, which corresponds to the lowest AIC, is highlighted. In nine out of twelve networks, the definition `pub-in` of the similarities is selected. For two patent networks, PAT 424 and PAT 703, `aut-in` is selected. In the case of RMP no conclusive outcome is reached: although `aut-out` has the lowest AIC (only marginally lower than `pub-out`), the parameter $\beta_1^{(\Omega)}$ is not significant. Summarising these outcomes, we find that in the overwhelming majority of cases, the asymmetric co-citation similarity is better than the bibliographic coupling in explaining the observed citations among authors. In most of the cases, similarity based on the citing publications is the best,

Table 5.1: The multiplex network regression for the citations among the 200 most cited authors on four definitions of author similarity. The networks correspond to four journals in the INPIRE data set (cf. Section 2.3).

Similarity	$\beta_0^{(\Omega)}$	$\beta_1^{(\Omega)}$	AIC
JHEP			
aut-in	5.603 ± 27.529	2.176 ± 0.011 ***	181727.3
aut-out	5.620 ± 26.537	2.129 ± 0.012 ***	189049.1
pub-in	2.686 ± 0.031 ***	1.124 ± 0.004 ***	118326.0
pub-out	2.983 ± 0.042 ***	1.292 ± 0.004 ***	122362.8
PR-HEP			
aut-in	9.483 ± 44.565	2.552 ± 0.016 ***	59196.46
aut-out	9.486 ± 44.658	2.362 ± 0.014 ***	64481.98
pub-in	3.499 ± 0.034 ***	1.166 ± 0.007 ***	41984.61
pub-out	3.587 ± 0.037 ***	1.135 ± 0.006 ***	48575.00
Phys. Lett.			
aut-in	4.407 ± 0.307 ***	1.985 ± 0.027 ***	17793.50
aut-out	2.705 ± 0.045 ***	1.538 ± 0.021 ***	20228.22
pub-in	2.571 ± 0.031 ***	1.038 ± 0.016 ***	16283.84
pub-out	2.303 ± 0.023 ***	1.003 ± 0.015 ***	18756.84
Nuc. Phys.			
aut-in	4.356 ± 0.251 ***	2.062 ± 0.017 ***	45769.73
aut-out	4.501 ± 0.307 ***	1.928 ± 0.015 ***	49974.41
pub-in	2.952 ± 0.033 ***	1.162 ± 0.010 ***	37863.43
pub-out	2.974 ± 0.034 ***	1.196 ± 0.009 ***	41902.91

while in two classes of patents, it is the one based on the citing authors. In a related study, Boyack and Klavans compare bibliographic coupling and co-citation similarity between publications in terms of their accuracy in clustering biomedical research, with the ground-truth based on the textual similarity [23]. Boyack and Klavans find that bibliographic coupling slightly outperforms co-citation similarity, which contrasts our finding. This mismatch, however, may be the result of different performance criteria: for us, it is explaining the observed citations, and for Boyack and Klavans, it is mapping science by means of clustering.

5.2.1 Friends and foes

Having selected the best formulation of the generalised hypergeometric ensemble, we proceed to the inference of signed relations between top cited authors. To recall, we have used a temporal order preserving definition of the combinatorial component in the ensemble, i.e., we substituted the matrix $\boldsymbol{\Xi}$ of the configuration model by Eq. (5.1). We have also compared four different similarity definitions and selected the one that best describes the observed network of citations between the top cited authors. Figures 5.6(a) and 5.6(b) show the matrices $\boldsymbol{\Xi}$ and $\boldsymbol{\Omega}$ used in the ensemble for the network of PR-HEP. The rows and columns of the presented matrices are ordered according to average linkage hierarchical clustering on the $\boldsymbol{\Omega}$ matrix[1], which results in topically closer authors also being positioned closer in the matrix [97, 99]. We apply the measure of signed deviations Φ defined in Eq. (4.6) to the marginal distributions of the hypergeometric ensemble (cf. Eq. (3.6)) to obtain the matrix of signed relations shown in Fig. 5.6(d). For comparison, we also compute signed relations based on the unbiased ensemble, i.e., the ensemble that disregards the edge propensities $\boldsymbol{\Omega}$, shown in Fig. 5.6(c). The comparison of the two signed matrices highlights how instrumental it is to account for the topical similarities of the authors when modelling the citations among them. The unbiased ensemble identifies most of the author pairs as under-citing each other, while for topically closer authors (pairs close to the diagonal of the matrices) who do cite each other, it mostly shows strong positive relation indicating over-citations. Instead, the ensemble that accounts for the similarities results in neutral ($\Phi \approx 0$) relations among the authors who are topically further from each other. For the authors that are topically close to each other, it results in non-trivial values of signed relations. Similarly to Fig. 5.6(d), Fig. B.1 shows the matrices $\boldsymbol{\Phi}$ for all twelve studied empirical networks.

Reciprocity in citations Similar to the analysis in Section 4.2.2, let us now compute the reciprocity among the top cited authors based on the inferred signed relations $\boldsymbol{\Phi}$. First, we compute the number of author pairs in the network that reciprocate or anti-reciprocate. A pair of authors reciprocates positively (++) if both authors over-cite each other, they reciprocate negatively if they both under-cite each other (−−), and they anti-reciprocate (−+) if the relations in the opposite direction have different signs. We have seen that the fractions of these three types of reciprocation depend strongly

[1]Hierarchical clustering is based on distances, which must be symmetric, so we applied it to the symmetric matrix with elements $\max(\Omega_{ij}, \Omega_{ji})$.

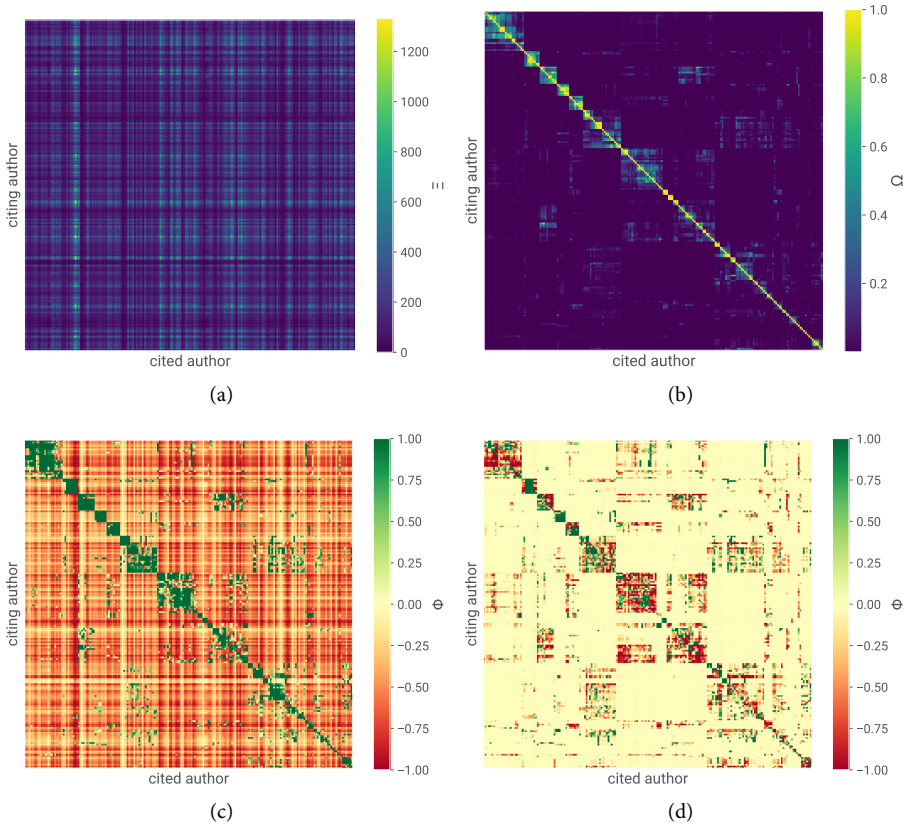


Figure 5.6: Matrices describing the 200 most cited authors in PR-HEP: (a) $\boldsymbol{\Xi}$ showing the total number of temporally possible citations for each pair of authors, (b) edge propensities $\boldsymbol{\Omega}$ based on `pub-in` definition of author similarities, (c) signed relations $\boldsymbol{\Phi}$ based on the unbiased hypergeometric ensemble with possibility matrix $\boldsymbol{\Xi}$, and (d) signed relations $\boldsymbol{\Phi}$ based on the generalised hypergeometric ensemble with possibility matrix $\boldsymbol{\Xi}$ and edge propensities $\boldsymbol{\Omega}$. The order of rows and columns corresponds to the hierarchical clustering performed on $\boldsymbol{\Omega}$.

on the threshold above which the signed relations are considered. Here we choose a threshold $\alpha = 0.5$, such that only the pairs of authors for whom both $|\Phi_{ij}| > \alpha$ and $|\Phi_{ji}| > \alpha$ are counted. The choice of the threshold approximately corresponds to the upper and lower quartiles of the signed relations. Table 5.2 shows the outcomes for the twelve empirical networks. The median fractions of the three types of reciprocation are 0.25 for $(++)$, 0.38 for $(-+)$ and 0.37 for $(--)$. For the following networks, these fractions deviate notably from the median (not accounting for RMP, for which there is

Table 5.2: Reciprocity in signed relations. The count and fraction of reciprocating (++ and −−) and anti-reciprocating (−+) pairs of authors for whom both $|\Phi_{ij}| > \alpha$ and $|\Phi_{ji}| > \alpha$ with $\alpha = 0.5$ (cf. Table 4.2). Two reciprocity measures for each network are presented, ρ defined in Eq. (4.13) and ρ^{sign} defined in Eq. (4.14)

Network	++	−+	−−	ρ	ρ^{sign}
PR	126 (0.20)	374 (0.59)	131 (0.21)	-0.068	0.832
PRA	403 (0.25)	790 (0.48)	439 (0.27)	-0.028	0.616
PRC	541 (0.13)	581 (0.14)	3040 (0.73)	0.646	0.860
PRE	222 (0.27)	355 (0.44)	231 (0.29)	0.039	0.818
RMP	5 (1.00)	0 (0.00)	0 (0.00)	0.015	0.078
JHEP	782 (0.21)	1636 (0.43)	1393 (0.37)	0.088	0.394
PR-HEP	163 (0.14)	452 (0.40)	517 (0.46)	0.150	0.835
Phys. Lett.	117 (0.37)	116 (0.37)	80 (0.26)	0.072	0.804
Nuc. Phys.	326 (0.35)	361 (0.39)	245 (0.26)	0.131	0.817
PAT 320	78 (0.16)	130 (0.27)	280 (0.57)	0.297	0.599
PAT 424	35 (0.56)	3 (0.05)	25 (0.40)	0.514	0.563
PAT 703	204 (0.57)	21 (0.06)	132 (0.37)	0.628	0.625

not enough data). In PRC, 73% of the identified pairs have a mutually negative relation. In patent networks PAT 424 and PAT 703, only 5% and 6% of the pairs anti-reciprocate. PAT 424 and PAT 703 networks are also the only ones in which more than half of the identified pairs exhibit positive reciprocation.

Next, we compute the aggregate reciprocity values for the networks according to the two definitions ρ and ρ^{sign} (cf. Eqs. (4.13) and (4.14)) introduced in Section 4.2.2. The first definition, ρ, corresponding to the definition in [77] accounts for both the absolute value and the sign of the relations Φ_{ij} between all pairs of authors i and j. The second definition, ρ^{sign}, defined in Eq. (4.14) only accounts for the signs of the relations. For all but two networks, the overall reciprocity among authors is positive according to both definitions, meaning that the pairs of authors tend to reciprocate more often than they anti-reciprocate. The exceptions are PR with $\rho = -0.068$ and PRA with $\rho = -0.028$. We find that the magnitude of reciprocity ρ is very small for most of the networks. The exceptions are the three patent networks and one scientific journal, PRC. In the patent networks, the reciprocity is considerably lower for the class PAT 320, which focuses on software solutions, as opposed to the physical R&D focused PAT 424 and PAT 703.

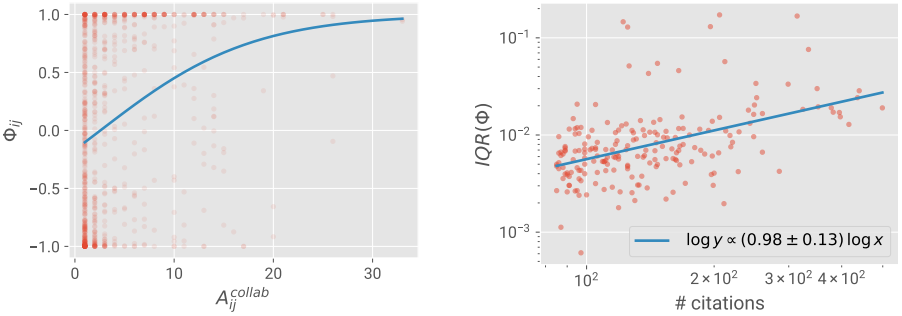


Figure 5.7: (Left) Logistic regression for the sign of Φ_{ij} on the number of collaborations between authors i and j (cf. Eq. (5.11)) and (right) linear regression for inter-quartile range $IQR_j(\Phi_{ji})$ of the signed relations towards author i on the total number of citations of i for the 200 most cited authors in PR-HEP.

Signed relation and coauthorship Let us investigate the relationship between the signed values Φ_{ij} and the number of coauthorships A_{ij}^{aa} between authors i and j. We would expect that frequent collaborators would have more positive relations, i.e., they would over-cite each other. To find this relationship, we perform a logistic regression [93] for the sign of Φ_{ij} on the number of coauthorships A_{ij}^{aa} as follows

$$\Pr(\Phi_{ij} \geq 0) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 A_{ij}^{aa})}. \tag{5.11}$$

In the left panel of Fig. 5.7, we show the result for the 200 most cited authors in PR-HEP. We see that, indeed, frequent coauthors tend to over-cite each other.

However, performing the same analysis for all twelve networks shows that PR-HEP is rather an exception. Table 5.3 reports the outcomes of the logistic regression of all networks, along with the R_{MF}^2 McFadden pseudo-R^2, which shows the quality of the fit [121]. Higher values of R_{MF}^2 indicate better regression models, with values 0.2–0.4 being considered a very good fit. We find that only for four networks, the dependence of the sign of Φ_{ij} on A_{ij}^{aa} (parameter β_1) is significant. For one of these four, PAT 703, the dependence is negative. And according to the values of R_{MF}^2, the logistic regression is overall a good model only for PRC. In Fig. B.2, we show the plots of the logistic regression for all twelve collaborative knowledge networks. Visual inspection of the plots confirms that we do not find compelling evidence for collaborative knowledge networks in general that frequent coauthors tend to over-cite each other.

Table 5.3: Logistic regression for the sign of Φ_{ij} on the number of collaborations between authors i and j (cf. Eq. (5.11)) in twelve collaborative knowledge networks. R^2_{MF} is the McFadden pseudo-R^2.

Network	$\beta_0^{\log}$	$\beta_1^{\log}$	R^2_{MF}
JHEP	0.219 ± 0.084 **	0.152 ± 0.031 ***	0.017
PR-HEP	-0.336 ± 0.096 ***	0.131 ± 0.020 ***	0.050
Phys. Lett.	0.529 ± 0.132 ***	-0.007 ± 0.029	-0.001
Nuc. Phys.	0.594 ± 0.108 ***	-0.016 ± 0.025	-0.000
PR	0.240 ± 0.141 .	0.059 ± 0.035 .	0.008
PRA	0.181 ± 0.080 *	0.024 ± 0.013 .	0.004
PRC	-0.683 ± 0.083 ***	0.330 ± 0.019 ***	0.222
PRE	0.556 ± 0.109 ***	0.003 ± 0.017	-0.000
RMP	-2.251 ± 3.793	2.326 ± 3.790	0.012
PAT 320	0.033 ± 0.144	0.043 ± 0.030	0.024
PAT 424	0.232 ± 0.247	0.013 ± 0.089	0.003
PAT 703	1.476 ± 0.145 ***	-0.075 ± 0.026 **	-0.018

Signed relation and author centrality Next, we analyse the relationship between the values of $\mathbf{\Phi}$ of a given author and the author's centrality in the collaborative knowledge network. We consider the degree centrality in the coauthorship network G^{aa} and in the projection of the citation network between publications G^{pc} onto a publication–author citation network. We compute the centralities based on the whole network in the selected ten year time window, i.e., the collaborative knowledge network of all articles published within the time window and their authors. Then, for a given author, the two centralities show her total number of unique co-authors and her total number of citations.

For an author i, we want to aggregate the behaviour of other authors towards her based on the values of Φ_{ji}. Specifically, we are interested in the *polarisation* of the behaviours towards her. We borrow the notion of polarisation from social and political sciences, where it measures the fragmentation and large-scale differences in opinions and ideological stances [75, 147]. Polarisation is also defined for networks of social interactions, under the assumption that differences in opinions and ideological stances have a behavioural component that influences the interactions between individuals. We follow the same assumption and measure the polarisation in interactions towards the author i as the *interquartile range (IQR)* of the vector $\Phi_{\cdot i} = \{\Phi_{1i}, \ldots, \Phi_{ni}\}$. That is, we take the difference between the 75th and 25th percentiles of the histogram of

the values of $\Phi_{\cdot i}$. From $\Phi_{ji} \in [-1, 1]$ follows that $IQR_j(\Phi_{ji}) \in [0, 2]$. The lowest value $IQR_j(\Phi_{ji}) = 0$ denotes that all values Φ_{ji} for a given i between the 75th and 25th percentiles are equal. We consider this case as the lowest polarisation, as it is a sign of a strong accordance among authors in their behaviour towards author i. The other extreme $IQR_j(\Phi_{ji}) = 2$ means that more than 25% of all authors j exhibit maximally negative relation $\Phi_{ji} = -1$ towards i, and more than 25% of all authors exhibit a maximally positive relation $\Phi_{ji} = 1$. Hence, in accordance with the accepted definitions of polarisation, this is the highest polarisation, as there are two groups with diametrically opposite behaviours.

We now perform a linear regression analysis (cf. Eq. (2.17) on Eq. (2.17)) for the logarithm of $IQR_j(\Phi_{ji})$ on the logarithm of the total number of citations and, separately, on the logarithm of the number of unique coauthors of author i. In the right panel of Fig. 5.7, we show the result of the regression on the the number of citations for the 200 top authors in PR-HEP and in Table 5.4, we provide the outcomes for all twelve empirical collaborative knowledge networks, along with the coefficient of determination of the regression model (cf. Eq. (2.18) on Eq. (2.18)). We find that for ten out of twelve networks, there is a significant positive dependence between the polarisation towards an author and the number of citations she received (parameter β_1^{cit}). The exceptions are RMP and PAT 424. The values of R^2 for the significant models are in the range of 0.06–0.3, which indicates that the number of citations is not enough to predict the polarisation towards an author. Still, up to 30% of the variance in the polarisation is explained by the number of citations. For the relation between polarisation and the number of coauthors, we find that for eight out of twelve networks there is a significant dependence expressed by the significance of the parameter β_1^{col}. These networks coincide with the ones for which the total number of an author's citations was significant in explaining the polarisation, with the exception of PR and PRC, for which the number of citations is significant, but the number of coauthors is not. We also find that the dependence of the polarisation on the number of citations is stronger in all significant cases, as for all studied networks, $\beta_1^{\text{cit}} > \beta_1^{\text{col}}$.

How can we interpret these results? In the ideal case, we could say that the more central, or prominent, an author is, the stronger and more diverse are the social biases of other authors towards her, expressed by means of over-citations and under-citations. For this conclusion to be valid, there is a set of conditions that must be met. First, in order to credit the identified over-citations and under-citations to social aspects, we need to make sure that the underlying null model accounts for all non-social, scientifically justified aspects that may influence the formation of citations. Second,

Table 5.4: Linear regression analysis for the interquartile range $IQR_j(\Phi_{ji})$ of the signed relations towards author i on (left) the total number of citations, and (right) on the number of unique coauthors of author i.

Network	β_0^{cit}	β_1^{cit}	R^2	β_0^{col}	β_1^{col}	R^2
PR	-3.685	0.800 ± 0.110 ***	0.214	-1.966	0.174 ± 0.089 .	0.024
PRA	-4.417	1.361 ± 0.239 ***	0.141	-1.770	0.935 ± 0.116 ***	0.255
PRC	-4.313	1.404 ± 0.373 ***	0.067	-1.017	-0.127 ± 0.163	0.003
PRE	-4.140	1.068 ± 0.131 ***	0.258	-2.034	0.275 ± 0.088 **	0.051
RMP	-1.557	0.027 ± 0.025	0.014	-1.553	0.034 ± 0.024	0.027
JHEP	-1.580	0.482 ± 0.086 ***	0.139	-0.708	0.469 ± 0.062 ***	0.228
PR-HEP	-4.212	0.982 ± 0.133 ***	0.215	-2.259	0.276 ± 0.092 **	0.047
Phys. Lett.	-3.152	0.728 ± 0.112 ***	0.188	-2.261	0.410 ± 0.108 ***	0.082
Nuc. Phys.	-3.611	0.954 ± 0.102 ***	0.310	-2.085	0.414 ± 0.093 ***	0.096
PAT 320	-3.397	1.033 ± 0.151 ***	0.216	-2.297	0.696 ± 0.130 ***	0.174
PAT 424	-9.321	0.130 ± 0.150	0.007	-9.324	0.164 ± 0.154	0.012
PAT 703	-2.997	0.571 ± 0.205 **	0.041	-2.613	0.531 ± 0.124 ***	0.103

we need to make sure that the dependence between polarisation towards an author and her centrality is not a result of confounding.

One specific way in which confounding threatens the validity of our results is the following. By construction, the signed measure of deviation Φ tends to be neutral for pairs of nodes with lower degrees (cf. Fig. 4.5), i.e., for authors with relatively low number of citations *from and to other top cited authors*. While the number of citations from other top authors is not directly related to the *total* number of citations in the *whole* collaborative knowledge network, the two can be correlated. We leave checking these threats to validity to future research.

5.2.2 The backbone of author–author citations

Lastly, let us look at the backbone of the network of citations among the most cited authors. Following the approach presented in Section 3.3, we identify as the backbone of the network all the pairs of nodes that cite each other significantly more frequently than predicted by an ensemble of random networks. We do not show the backbone based on percentile filtering, as in Chapter 3. Instead, we directly apply the filtering based on the inferred signed relations, as presented in Section 4.2. As in Chapter 4, at first we look at the backbone based on the unbiased hypergeometric ensemble. Here

we use the ensemble with order preserving configuration defined in Eq. (5.1). Hence, for the network of the 200 most cited authors in PR-HEP, the filtering is performed on the signed matrix shown in Fig. 5.6(c), to which we refer as $\mathbf{\Phi}^{(0)}$. That is, we include in the backbone all pairs of nodes for which $\Phi_{ij}^{(0)} > 1 - \alpha$. Choosing the threshold $\alpha = 0.01$, we obtain the backbone shown in Fig. 5.8(b). The whole citation network of the 200 most cited authors is shown in Fig. 5.8(a). In all networks in Fig. 5.8, the multiple citations between a given pair of authors (the multi-edge) are visualised as one edge with the width proportional to the number of citation between the pair. While we can see two pronounced communities in the whole network, the backbone exposes more intricate sub-community structure within the two large communities. In the backbone, there are 1560 (directed) connected pairs of authors with a total of 27122 individual citation edges, compared to 3904 connected pairs with 31402 individual citation edges in the whole network. That is 86.4% of all citations among 40.0% of all pairs that cite each other.

Inferring an alternative network backbone based on the biased hypergeometric ensemble that accounts for the topical similarity between authors (co-citation similarity based on the citing publications, cf. Table 5.1) leads to the network shown in Fig. 5.8(c). It has 479 connected pairs that have 12737 individual citations among them, which is 12.3% of the connected pairs (significantly more than expected by chance at $\alpha = 0.01$) and 40.6% of all citations in the whole network. Compared to Fig. 5.8(b), the sub-community structure has disappeared and even the two large communities became less pronounced, as there are now more nodes in between them. This is an expected outcome. The backbone based on the unbiased ensemble mainly highlights the topical community structure, which is also evident from the topical similarities among the authors (cf. Figs. 5.6(b) and 5.6(c)). Instead, the backbone based on the ensemble that encodes the topical similarities exposes the pairs of authors that over-cite each other *beyond* what is predicted by the topical similarities. The interpretation of this phenomenon is not simple. As discussed earlier, we cannot safely say that all pairs in the backbone shown in Fig. 5.8(c) over-cite each other due to non-scientific social aspects, unless we make sure that all scientifically justifiable aspects are incorporated in the ensemble that serves as the null model. However, given that we account for a major factor that influences the citations—the topical similarity—we can argue that pairs connected in the network of Fig. 5.8(c) comprise all the *candidates* for whom positive social biases may be at play (similarly, the candidates for possible negative social biases can be identified by filtering the significant under-citations according to $\Phi_{ij}^{(\Omega)} < -1 + \alpha$). For example, we see a clique of five nodes circled in red in

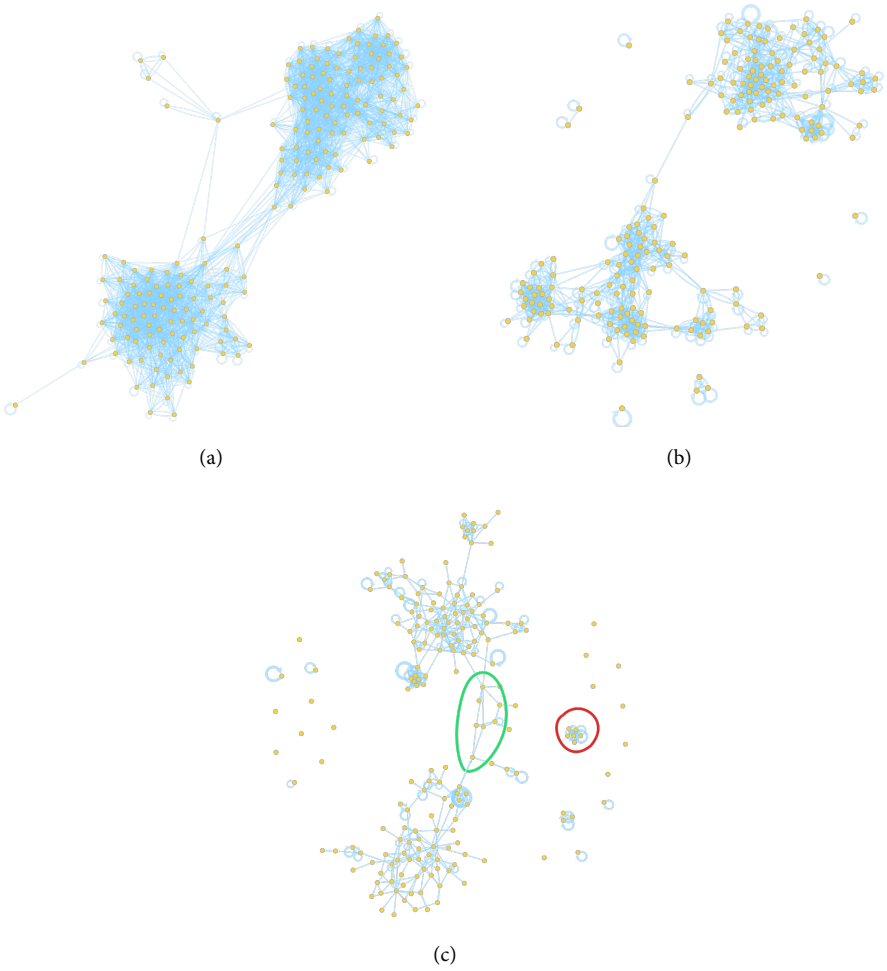


Figure 5.8: Citations among the 200 most cited authors in PR-HEP: (a) the full network, (b) the network filtered by the condition $\Phi_{ij}^{(0)} > 1 - \alpha$ and (c) the network filtered by the condition $\Phi_{ij}^{(\Omega)} > 1 - \alpha$. In both (a) and (b), $\alpha = 0.01$. Edge widths are proportional to the number of citations between the corresponding pair of authors.

Fig. 5.8(c), which are disconnected from the rest of the network. We can speculate that these authors might constitute a *citation cartel* [67]. On the other hand, we see a set of authors in between the two communities (circled in green), which are not seen in the whole network visualisation or in the backbone from the unbiased ensemble. It is plausible to assume that these authors pioneer an interdisciplinary topic that

lays between the main topics of the two large communities and that they cite each other only due to scientifically justifiable reasons. Then, our method identifies them as over-citing each other because their interdisciplinary research is not *yet* reflected in the co-citation similarity between them. In order to turn these speculations to sound scientific outcomes, we need to further refine the methodology presented in this chapter in future research. For instance, to identify citation cartels, we can also add keyword-based similarity in the null model, from which it will follow that over-citing authors do so beyond the prediction based on the community practices (co-citation similarity) and on the similarity of the content between their citing and cited publications. The comparison of the outcomes between network-based and keyword-based similarities can also help to identify the interdisciplinary "pioneers".

5.3 Conclusion

In this chapter, we have adapted the methodology developed in Chapters 3 and 4 for the network of citations among authors. We have done so by replacing the combinatorial component of the generalised hypergeometric ensemble based on the configuration model by a model that respects the temporal order of citations. We have encoded the topical similarities between authors in the ensemble by means of multilayer network regression. This allowed us to compare different formulations of author similarity in terms of their power to explain the empirical citation network. With the best similarity formulation chosen, we inferred signed relations between the top cited authors in each studied collaborative knowledge network. That is, to interpret the relations between individual scientists, we looked at them through the lens of their scientific community as a connected whole. These signed relations represent over-citations and under-citations among these authors.

The analysis of the inferred signed relations among the authors revealed the following findings. Contrary to our prior expectation, it is only in few of the studied empirical data sets we have found a significant difference between the signed relations among researchers who have previously co-authored and those who have not. Our prior expectation was based on the hypothesis that coauthors, who are acquainted, (i) are prone to selection bias and (ii) have a positive predisposition towards each other, both of which would be reflected in their citation behaviour. Instead, coauthors on average cite each other "fairly", as expected from their topical similarities.

However, we found that more prominent authors—measured by the number of cita-

tions and the number of coauthors they have—tend to be more polarising in the sense that there are more diverse expressions of over-citation and under-citation from other authors towards them. At this stage, we did not exclude confounding effects that could undermine this finding, leaving that to future research. If this finding is confirmed, we can argue that it highlights the increasing competition among authors with the increase of their prominence. To address the issue of confounding, we could build a baseline model, either a random network ensemble or an agent-based model, that replicates the empirical numbers of publications and citations of the authors, as well as their coauthorship relations. Then, this baseline model would show how much of the observed tendencies could be expected at random.

We also identified that the authors tend to reciprocate in all studied empirical networks, meaning that a pair of authors tends to have similar citation behaviour towards each other—either mutually over-citing or mutually under-citing each other. This tendency is very mild in most of the scientific networks, but is quite strong in the patent networks. In patents, the reciprocity is much higher in classes that are related to hardware and pharmaceutical inventions, compared to the network related to software solutions. As mentioned in Chapter 1, it is well established that humans tend to reciprocate positive and negative relations. Thus, the fact that the inferred over-citations and under-citations tend to be reciprocal, lends credibility to the claim that our method exposes traces of social biases.

Lastly, we illustrated how the structure of a scientific community can be visualised by means of network backbone inference based on the order preserving null model of citations. Moreover, we have outlined how the inspection of the backbone based on the null model that incorporates topical similarities can help in identifying citation cartels and interdisciplinary pioneers.

We believe that the presented approach to analysing relations between authors has an applied value. It can facilitate the decision making in academia by, for instance, providing additional information to journal editors or funding agencies about author relations when selecting peers for manuscript reviews or grant proposal evaluations. We will further discuss these implications in Chapter 8.

Part II

Dynamics

Life is growth. If we stop growing, technically and spiritually, we are as good as dead.

Morihei Ueshiba

Chapter 6

Growth of collaborative knowledge networks

Summary

We study the evolution of collaborative knowledge networks by means of generative growth models. We formulate coupled models for citation growth that account not only for the growth history of the citation network itself (citation component), but also for network position of the authors of the cited artifacts (social component). By fitting growth models based on individual growth events, instead of aggregate properties of the final networks [124], we further this approach by estimating the statistical errors of the growth model parameters, which are commonly neglected. We show that efficient model fitting can be performed by sampling growth events uniformly through the evolution of the network. For twelve empirical collaborative knowledge networks, we compare additive and multiplicative forms of coupling between various citation and social components. We find that in most of the cases, a model with coupling describes the observed network best. Moreover, we find that including a social component that characterises all of the authors of an artifact, leads to a better model, compared to one that only characterises the single most prominent author. While we only investigate the coupled growth of citations, we also outline a more comprehensive model for the simultaneous growth of both citations and authorship relations.

This chapter has been written specifically for this dissertation.

In the second part of this dissertation, we address the dynamics of collaborative knowledge networks. As already discussed in Chapter 1, the collaborative knowledge networks grow over time. New knowledge artifacts—scientific publications, or patents—are constantly added to the network, along with their citation edges to existing artifacts and the authorship edges to their authors (or inventors). Some of the authorship edges are drawn to incumbent authors who are already in the network due to previously authored knowledge artifacts. Others are drawn to newcomer authors, who are added to the network at the same time as the corresponding knowledge artifact. Figure 6.1 shows the growth of a small part of a collaborative knowledge network in High-Energy Physics (INSPIRE data set). It starts with the leftmost snapshot and from there, each successive snapshot shows the state of the network at a later time step.

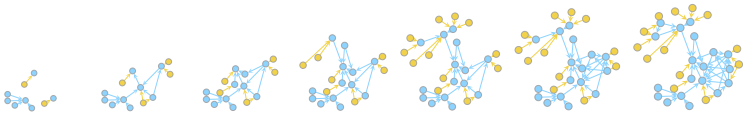


Figure 6.1: Growth of a collaborative knowledge network.

The formation of edges in the network over time is driven and affected by various mechanisms. As discussed in Chapter 5, topical similarity plays an important role in the formation of citations, with the tendency that more citations are drawn between the artifacts that are topically more similar. There is also evidence of preferential attachment in citations, according to which artifacts that already have a large number of citations tend to attract even more citations [9, 157, 159]. Another important mechanism is the *ageing* of knowledge artifacts—also referred to as *relevance decay* or *novelty decay*—with older artifacts attracting less citations over time [81, 125, 149, 211]. For authors, new coauthorship edges tend to be formed between the authors who have previously co-authored artifacts [30, 74]. Preferential attachment is at play also in coauthorship formation [133, 195].

The aforementioned mechanisms take into account and affect a single (either citation- or coauthorship-) layer in collaborative knowledge networks. So far, very limited research has been dedicated to the analysis of inter-layer mechanisms affecting the formation of such networks. However, as early as 1968 in his seminal work Merton discussed various mechanisms of how characteristics of authors (prominence, academic awards) may be influencing the recognition of their publications (in terms of citations) [126]. In particular, he raised the question of whether the publications by

better known researchers get more and faster recognition in the community in terms of citations. It has been recently shown that, indeed, the centrality of authors in the coauthorship network prior to a publication affects the number of citations that the publication will receive [165]. Similarly, it has been shown that authors with many citations tend to attract more coauthors in the future [133]. In this part of the dissertation, we aim to advance the line of research that investigates the inter-dependencies of the dynamics of the different layers in collaborative knowledge networks.

In this chapter, we investigate the *growth* of collaborative knowledge networks over time. Specifically, we focus on the coupling between the growth mechanisms of citations and coauthorship layers.

6.1 Modelling approach

In Part I, we investigated the structure of collaborative knowledge networks using statistical models that aimed to explain the observed networks. As such, the assumptions about the data generating processes were quite general and not limited to the collaborative knowledge networks. Namely, we assumed a biased urn model underlying the formation of the aggregate network that can be used in any setting where repeated interactions between nodes are observed. In contrast, in this part, and in this chapter specifically, we investigate the *dynamical mechanisms* underlying the formation of collaborative knowledge networks. Here, too, we employ a generative statistical modelling approach, but such that specific dynamical rules of the network *evolution* are the focus.

The overwhelming majority of studies on growing networks have focused on fitting models of growth to *macroscopic* patterns of networks, such as the degree distribution or community structure [14, 128, 133, 143, 176, 202]. Following this approach in our previous research, we used *master equations* from statistical physics to study the degree distribution of co-evolving networks that follow the preferential attachment and clique formation mechanisms [133].

In this chapter, however, we evaluate and compare different growth mechanisms on a *microscopic level* [109, 124, 127]. Specifically, we evaluate how well a model can explain the sequence of individual growth events in a network, instead of how well aggregate network measures resulting from the models fit the observed data. This way we are able to draw conclusions about the models directly with respect to growth process,

and not indirectly from statistical properties of the resulting aggregate network.

6.1.1 Coupled growth models

In this section, we formulate *coupled* growth models for the citation network between knowledge artifacts and for the corresponding coauthorship network between the contributors to these artifacts. Citation networks and their growth are extensively studied in isolation, i.e., without accounting for the influences from the coauthorship network [9, 105, 149, 159, 204, 205, 206]. As a result, there is an agreement in the research community about general mechanisms involved in the growth of citations, such as (i) *fitness* of the publication that influences how attractive the publication is for citations [204], (ii) preferential attachment [2, 9, 159, 205, 206], (iii) decay of the relevance of knowledge artifacts over time [79, 105, 149, 206].

In Section 2.4.1, we discussed a class of generative network growth models (cf. Eqs. (2.12) to (2.14)) that build on the principle of proportional growth and are designed to model the above-mentioned mechanisms. Starting with these models intended for single-layer networks, we define growth models for the coupled evolution of two layers of collaborative knowledge networks.

We study two general forms of coupling between two network layers, *additive* and *multiplicative*. Without loss of generality, we focus on the formation of new citations between knowledge artifacts, dependent on (i) the existing citation network and (ii) a projection layer of the collaborative knowledge network on authors. We refer to the first model component that is based on the citation network as `cit`, and to the second model component as `soc`, for "social", as it is based on a network of authors. We write the additive form of coupling as

$$P_j(t;\beta) = \beta^{+} P_j^{cit}(t) + (1-\beta^{+}) P_j^{soc}(t), \tag{6.1}$$

which means the probability $P_j(t)$ to connect to artifact node j at time t is the weighted mixture of two components defined on the citation layer, P_j^{cit}, and on the network layer P_j^{soc} which has the authors as nodes. The parameter β^{+} is called *mixture weight*. Such mixture models are used when the statistical population is known to comprise sub-populations that are described by different probability distributions. The mixture weight is, then, determined by the relative size of each sub-population. In our case, the mixture formulation can be interpreted as if the node is chosen due to *either* one

process, *or* the other, with the respective weights.

Before formulating the second, multiplicative, form of coupling, we introduce the notion of *odds*. As explained in Chapter 3, the odds of an event reflect its relative likelihood. The odds are proportional to the probability up to a constant factor, such that

$$P_j^{cit/soc}(t) = \frac{W_j^{cit/soc}(t)}{\sum_l W_l^{cit/soc}(t)}, \tag{6.2}$$

where $W_j^{cit/soc}(t)$ is the odds to select the node j at time t. For instance, the linear preferential attachment (cf. Eq. (2.11)) then writes as $W_j = k_j$, i.e., the degree of the nodes precisely sets the odds to select that node.

Using Eq. (6.2), we can write the multiplicative coupling form as

$$P_j(t;\beta) = \frac{W_j^{cit}\,[W_j^{soc}(t)]^{\beta^*}}{\sum_l W_l^{cit}\,[W_l^{soc}(t)]^{\beta^*}}. \tag{6.3}$$

We can interpret this coupling form as follows: the social component of growth *scales* the effect of the citation component. That is, we see the growth component defined on the citation layer as the main, baseline effect, while the social component biases it. The formulation is similar to the notion of propensities introduced in Chapter 3. The meaning of the parameter β^* also differs from β^+ in Eq. (6.1) for additive coupling: it denotes the *strength* of the influence that the social component has on the growth of citations.

6.1.2 Citation component

With the two general coupling forms introduced, we proceed with the specific formulations of the two constituent components $P_j^{cit}(t)$ and $P_j^{soc}(t)$. We consider three candidates for the citation component $P_j^{cit}(t)$. We write these using the notation introduced in Section 2.4.1.

PA: Linear preferential attachment with constant additive term,

$$W_j^{cit}(t) = k_j^{\text{in}}(t) + \alpha^+, \tag{6.4}$$

where k_j^{in} denotes the in-degree of the knowledge artifact j in the citation network $G^{pc}(t)$ (cf. Section 2.2), i.e., the number of citations of j at time t. The parameter α^+ can assume any positive non-zero value, $\alpha \in (0, \infty)$. It cannot be zero as, in that case, nodes with $k_j^{\text{in}} = 0$ would never acquire citation edges according to Eq. (6.4).

PA-RD: Linear preferential attachment with constant additive term scaled by relevance decay,

$$W_j^{cit}(t) = (k_j^{\text{in}}(t) + \alpha^+)e^{-\frac{t-t_j}{\tau}}, \tag{6.5}$$

where t_j denotes the time at which the node j was added to the network, thus the odds for node j to be cited at time t now also depend on the age of the node. For simplicity, we will measure the time in terms of the number of knowledge artifact nodes in the network. This means then $t_j = j$ if the nodes are labelled according to the order in which they are added to the network. The formulation in Eq. (6.4) is a special case of Eq. (2.14). The parameter τ determines the characteristic time of relevance decay and can assume any positive non-zero value, $\tau \in (0, \infty)$.

PA-NL-RD: Non-linear preferential attachment scaled by relevance decay,

$$W_j^{cit}(t) = (k_j^{\text{in}}(t) + 1)^{\alpha^*} e^{-\frac{t-t_j}{\tau}}. \tag{6.6}$$

Similar to the PA model, we need to offset the degree to allow the zero-degree nodes to attract their first edge. We choose to fix the offset to one to keep the model complexity of the citation component the same in all definitions. That is, we still have one parameter, α^*, controlling the degree related preference. In this case, it amplifies or weakens the effect of the number of previous citations.

6.1.3 Social component

Let us now proceed with the formulation of a candidate set of social components $W^{soc}(t)$. We will investigate three conceptually different formulations.

NAUT: First, we will study the effect of the *team size* on the citation dynamics. The team size is the number of authors that wrote a given knowledge artifact. We ask whether there is preference to cite artifacts that are made by larger teams [126].

The hypothesis behind this is the following: the larger the group of people that contributes to a knowledge artifact, the more cumulative effort is put into it, which ideally translates into higher scientific impact of the artifact and a higher number of citations. The team size of a knowledge artifact j is the in-degree $k_j^{(a)\text{in}}(t)$ in the bipartite network $G^a(t) = \left(V^a(t), V^p(t), E^a(t)\right)$ of the authorship edges between authors and publications (cf. Section 2.2):

$$W_j^{soc}(t) = k_j^{(a)\text{in}}(t). \tag{6.7}$$

NCOAUT: The second formulation for the social component accounts for the contributors' *centrality in the coauthorship network*. For simplicity, we will limit ourselves to the degree centrality. However, the method does not depend on this choice and can be expanded to other centrality measures. The centrality of an author in the coauthorship network is a proxy of how many people are acquainted with the author. We may assume that the direct academic acquaintances of an author have a better knowledge of the work of this author, meaning that the more acquaintances an author has, the higher is the visibility of her work in the community. We define two variants of this social component.

In the first variant `NCOAUT`, we aggregate the coauthors of all authors of a given knowledge artifact. We use the notion of paths (cf. Eq. (2.5)) on the bipartite network G^a corresponding to the authorship edges, in order to find the number of unique coauthors:

$$W_j^{soc}(t) = \left|\{v \mid \exists \pi_{3,jv}(t)\}\right|, \tag{6.8}$$

where we count the unique endpoints v of the paths $\pi_{3,jv}$ of length three. On the bipartite network, these paths starting at the knowledge artifact j necessarily end at a node corresponding to an author (artifact–author–artifact–author). Recall that in Section 2.1.2, we defined the paths $\pi_{\lambda,ij}$ as self-avoiding, meaning that in Eq. (6.8), we do not count in the authors of j. Note that the number resulting from Eq. (6.8) is different from the sum of the numbers of co-authors of each author of j, as the latter would count the common co-authors multiple times.

MAXCOAUT: In the second variant of the social component based on coauthorship centrality, we only account for the most central author [165]. For authors $l \in$

$V^a(t), (l, j) \in E^a(t)$ of the artifact j,

$$W_j^{soc}(t) = \max_l(\left|\{v \mid \exists \pi_{2,lv}(t)\}\right|), \tag{6.9}$$

where we count the number of unique coauthors of each author l by traversing the bipartite network $G^a(t)$ (2-paths author–artifact–author) and we take the largest count.

NPUB: The third formulation of the social component encodes the author' *academic experience* measured in terms of the number of previous publications. The hypothesis behind this formulation is that with growing experience of an author, each of her subsequent publications are able to attract more citations. As with the previous formulation, we define two variants.

In the first variant NPUB, for a knowledge artifact j, we count the number of distinct artifacts written previously by the authors of j.

$$W_j^{soc}(t) = \left|\{v \mid \exists \pi_{2,jv}(t)\}\right|. \tag{6.10}$$

As in Eq. (6.8), this is different from summing the number of publications of each of the authors, as doing so would count the publications co-authored by the authors of j multiple times.

MAXPUB: In the second variant of social component that aims to measure the academic experience, we only take the number of knowledge artifacts by one author j, who wrote the most of them. So, for authors $l \in V^a(t), (l, j) \in E^a(t)$ of the artifact j,

$$W_j^{soc}(t) = \max_l(k_l^{(a)\text{out}}(t)), \tag{6.11}$$

where the degree $k_l^{(a)\text{out}}$ of the author l in the authorship network $G^a(t)$ shows the number of artifacts written by l up to the time t.

The motivation to study the model variants MAXCOAUT and MAXPUB stems from the discussion about Matthew effect by Merton, where he conjectures that most credit for the publication is given to the most prominent author [126]. If these variants perform better than their counterparts that aggregate the network positions of all authors of an artifact, we can interpret that as support for Merton's conjecture.

6.2 Model fitting and selection

With candidate formulations for the two components in the model of coupled growth introduced, we want to compare them and select the model that best explains the growth of citations in empirical collaborative knowledge networks. In the following, we describe the model selection procedure. We use MLE approach to fit the parameters of a model to the data (cf. Section 2.4.2). The general form of the log-likelihood function provided in Eq. (2.15) writes our growth models as

$$\ln \mathcal{L}(\boldsymbol{\theta}; G) = \sum_{i=1}^{N} \sum_{(i,j)\in E^{pc}} \ln P_j(i; \boldsymbol{\theta}), \tag{6.12}$$

where G is the collaborative knowledge network at the end of the growth process, $\boldsymbol{\theta}$ is the vector of model parameters, the nodes $i \in V^p$ are enumerated according to when they are added to the network. Hence, the outer summation in Eq. (6.12) runs through the sequence of growth events in the network. The inner summation runs through the nodes $j \in V^p, j < i$ cited by the currently added node i. Finally, $P_j(i; \theta)$ is the probability to observe the citation (i, j) according to the considered model. Equation (6.12) is based on the assumption that edges are drawn independently.

Then, the model parameters $\hat{\boldsymbol{\theta}}$ that maximise the likelihood of the model given the data are found by solving

$$\left.\frac{\partial \ln \mathcal{L}}{\partial \boldsymbol{\theta}}\right|_{\hat{\boldsymbol{\theta}}} = 0. \tag{6.13}$$

An important feature of MLE method is that under certain mild conditions, the estimated parameters $\hat{\boldsymbol{\theta}}$ are normally distributed around the means that correspond to the true values of the parameters, with variance-covariances that can be expressed by the *Fisher information matrix* $\mathcal{I}(\hat{\boldsymbol{\theta}})$,

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) = \left[\mathcal{I}(\hat{\boldsymbol{\theta}})\right]^{-1}. \tag{6.14}$$

The Fisher information matrix is the expectation of the second derivatives of the log-likelihood function,

$$\mathcal{I}(\boldsymbol{\theta}) = -\,\mathrm{E}\left[\frac{\partial^2 \ell(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}}\right]. \tag{6.15}$$

In practice, the *observed Fisher information* matrix is used just by calculating the value

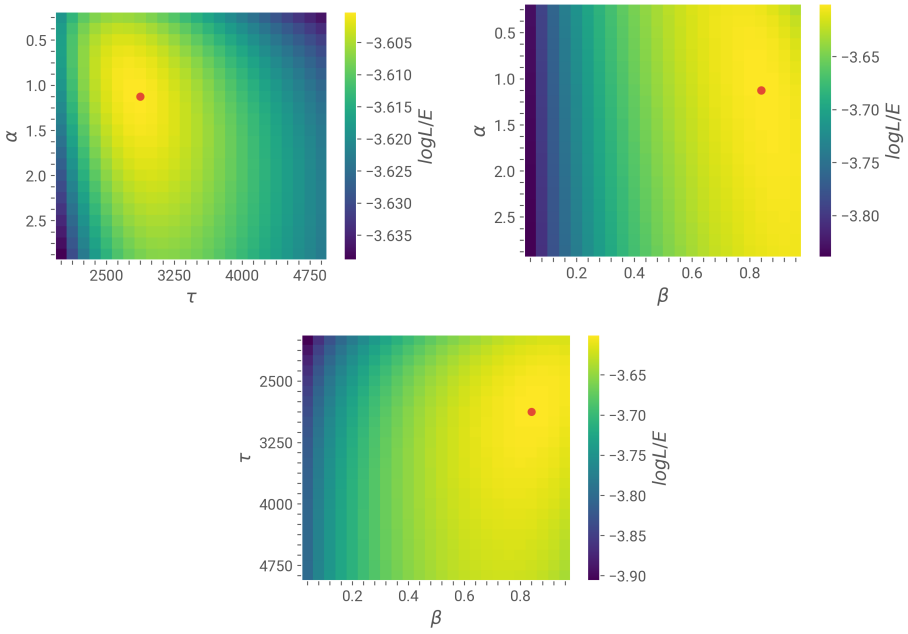


Figure 6.2: The log-likelihood function per edge for the model **PA-RD-NCOAUT** fitted to the collaborative knowledge network of JHEP. Each section plane of the shown two model parameters corresponds to the maximum likelihood value of the third parameter. The red dot shows the location of the maximum likelihood. It is the same in all three plots.

of the matrix at the parameter estimates, without applying the expectation operator. The variance of the parameter estimates is inversely proportional to the number of observations based on which the estimation is performed. This is because the Fisher information matrix can be expressed as $\mathcal{I}(\boldsymbol{\theta}) = N\mathcal{I}_1(\boldsymbol{\theta})$, where $\mathcal{I}_1(\boldsymbol{\theta})$ is the information matrix per observation.

For our problem of modelling the growth, we cannot analytically calculate the log-likelihood in Eq. (6.12) and its derivatives, due to different normalisation factors in each summand $P_j(t;\boldsymbol{\theta})$ (cf. Eq. (6.2)). Hence, we have to resort to numerical computation. However, because the formulated models are of well-behaving polynomial and exponential form, we know that the likelihood function is smooth and convex, at least locally [124]. To confirm that, in Fig. 6.2 we show the surface of the log-likelihood function for one model, which combines the citation component `PA-RD` and the social component `NCOAUT` according to the additive coupling form given by Eq. (6.1).

Instead of naively searching for the maximum of the likelihood function on a grid

of all possible parameters, we utilise its convex shape and apply a *greedy hill-climbing algorithm*. Such algorithms work as follows. Two values of the objective function (negative log-likelihood, in our case) are computed: for arbitrary initial parameters, and for parameters that differ slightly from the initial ones. The parameters corresponding to the lower value of the objective function are selected. In the next iteration, the value of the objective function of the selected parameters is compared to the value for slightly different parameters. The process is continued until convergence up to desired precision is achieved, at which point the resulting parameters are taken as the ones that minimise the objective function.

Some of the more advanced hill-climbing algorithms do not randomly choose the next parameter values in each iteration. Instead, they numerically estimate the gradient of the objective function at each iteration, in order to achieve the convergence in fewer iterations. To find the gradient, the algorithm computes the *Hessian matrix*, which for the log-likelihood function corresponds to the Fisher information matrix. That means that as a result of MLE, we estimate not only the parameter values but also their variances.

In other words, with the right choice of the optimisation algorithm, we are able to obtain the error estimates of the parameters. One suitable algorithm is *Broyden–Fletcher–Goldfarb–Shanno algorithm*. More specifically, we use the variant of the algorithm, `L-BFGS-B`, that allows to set boundary constraints on the parameters and that uses limited memory, making it more scalable [29, 99].

Once we obtain the maximum likelihood estimates of the parameters for all candidate models, we perform a model selection according to the relation likelihood of the models (cf. Eq. (2.21)) based on the Akaike Information Criterion (cf. Eq. (2.19)). We cannot perform a more principled likelihood ratio test, because the models we compare are not nested. We choose the Akaike Information Criterion over Bayesian Information Criterion for the same reason.

6.2.1 Sampling growth events

The independence between growth events, i.e., between newly created citation edges, which we assumed in our formulation of the growth models, allows us to implement a highly scalable computational procedure that, in principle, can be applied to very large networks. Because of this assumption, each summand of the log-likelihood

function in Eq. (6.12) corresponds to addition of a single edge. This means that we can split Eq. (6.12), corresponding to the whole growth history of the network, into log-likelihoods of arbitrary samples of growth events. Let us denote the growing collaborative knowledge network that has N consecutively added knowledge artifacts as $G_{1,N}$. Then, we can split the full series of growth events $1, \dots, N$ corresponding to the addition of new knowledge artifacts to the network $G_{1,N}$ into K consecutive subsets, with each subset adding $v = \lfloor N/K \rfloor$ knowledge artifacts. Each subsequent subset $l \in \{1, \dots, K\}$ starts with the network $G_{1,lv}$ and adds the next v knowledge artifacts. If we denote the sub-network that comprises the nodes and edges added in the l-th subset of events as $G_{(l-1)v,lv}$, we can rewrite Eq. (6.12) as

$$\ln \mathcal{L}(\boldsymbol{\theta}; G_{1,N}) = \sum_{l=1}^{K} \ln \mathcal{L}(\boldsymbol{\theta}; G_{(l-1)v,lv}), \tag{6.16}$$

where the log-likelihood for each subset l is given by

$$\ln \mathcal{L}(\boldsymbol{\theta}; G_{(l-1)v,lv}) = \sum_{i=(l-1)v}^{\max(lv,N)} \sum_{(i,j)\in E^{pc}} \ln P_j(i; \boldsymbol{\theta}). \tag{6.17}$$

Now, if we estimate the model likelihood for each sample $\mathcal{L}(\boldsymbol{\theta}; G_{(l-1)v,lv})$ individually, we obtain K parameter estimates $\hat{\boldsymbol{\theta}}_{\mathbf{1}}, \dots, \hat{\boldsymbol{\theta}}_{\mathbf{K}}$. As we discussed above, these parameter estimates are normally distributed with the mean at the true value, and their variance grows proportionally with decreasing number of events in the sample. This means that, in principle, one can adjust the sample size according to the desired precision level of the parameter values.

6.3 Outlook: simultaneous growth of citations and authorships

So far, we have only discussed the growth of the citation network between publications affected by the history of the whole collaborative knowledge network. However, the outlined procedure of model fitting and selection can be applied to a more comprehensive formulation that models the simultaneous evolution of both citations and authorship relations in the collaborative knowledge network. To obtain such a model, we need to add the component responsible for the growth of authorship relations to the already formulated component for the growth of citations (cf. Eqs. (6.1) and (6.3)).

The model component P^a responsible for the authorship edges can be defined in the same manner as we defined the component for citation formation. That is, we can write the probability P_j^a for an author $j \in V^a(t)$ to be connected to the node of the knowledge artifact that is newly added to the network at time t. The exact formulation for P_j^a can be made based on the existing literature on the formation of collaborative teams [30, 86, 128, 133]. The probability P_j^a of the node $j \in V^a(t)$ to be selected as the author at time t can depend, e.g., on the number of previous coauthors (coauthorship network G^{aa}); on the number of previously written publications (the bipartite network G^a); or on the number of previous citations of the author j [133]. That is, the model component for authorship formation can also implement a coupling between different layers of the growing collaborative knowledge network. The component P_j^a must also implement the addition of new authors to the network [74, 86].

If we denote the model component for the growth of citations, previously defined in Eqs. (6.1) and (6.3), as $P^{pc}(t;\boldsymbol{\theta}^{pc})$, the joint model of citation and authorship formation becomes

$$P_j(t;\boldsymbol{\theta}) = \begin{cases} P_j^{pc}(t;\boldsymbol{\theta}^{pc}) & \text{if} \quad j \in V^p(t), \\ P_j^a(t;\boldsymbol{\theta}^a) & \text{if} \quad j \in V^a(t). \end{cases}, \tag{6.18}$$

where $\boldsymbol{\theta}^{pc}$ are the parameters in the component of the citation formation and the $\boldsymbol{\theta}^a$ are the parameters in the component of authorship formation. That is, when an edge is formed at time t between the newly added knowledge artifact and another node j, that edge is modelled by P_j^{pc} if $j \in V^p(t)$ is a knowledge artifact, and by P_j^a if $j \in V^a(t)$ is an author.

So, multiple edges are created at each time step t that corresponds to the addition of a new knowledge artifact. All these edges have the newly added artifact on one end, and other artifacts (citation edge) or authors (authorship edge) on the other end. We do not have to explicitly model other types of edges, such as coauthorships or author–author citations, as these are projections of the two basic edge types that we model (cf. Sections 1.1 and 2.2). If we assume independence between these newly created edges, we can write the log-likelihood function of the model defined in Eq. (6.18) as

$$\ln \mathcal{L}(\boldsymbol{\theta}; G) = \sum_{i=1}^{N} \left[\sum_{(i,j)\in E^{pc}} \ln P_j^{pc}(\boldsymbol{\theta}^{pc}) + \sum_{(i,j)\in E^a} \ln P_j^a(\boldsymbol{\theta}^a) \right], \tag{6.19}$$

where $\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}^{pc} \\ \boldsymbol{\theta}^a \end{pmatrix}$ is the vector of all parameters of the model.

Once our model of simultaneous growth of the citations and authorships in a collaborative knowledge network is formulated according to Eqs. (6.18) and (6.19), we can straightforwardly apply the procedure described in Section 6.2 for model fitting and model selection.

6.4 Coupled growth of citations in empirical networks

In the following, we present the results of fitting the models defined in Section 6.1 to the twelve empirical networks introduced in Section 2.3. First, we will validate that the model parameters can be estimated based on a subset of the growth history of a collaborative knowledge network, as described in Section 6.2. Then, for one empirical network, we will discuss in detail the outcomes of different formulations of the model of citation growth. And last, we will present the best fitting model formulations for all studied networks.

Validation: sample size and temporal trends In Section 6.2.1, we described how the parameters of a growth model can be estimated based on a sample of growth events. In particular, we explained a procedure of dividing the whole history of the growth of a collaborative knowledge network into subsequent growth periods. Let us now apply that procedure to fit one model formulation to the data set of Nuc. Phys.. We choose the model with additive coupling (cf. Eq. (6.1)) between the citation component PA-RD defined in Eq. (6.5) and the social component NCOAUT defined in Eq. (6.8). In the following, we refer to this model as PA-RD-NCOAUT. Noting that the parameter estimation based on a sample is not limited to consecutive growth events (newly added artifacts), we first select, uniformly at random, 5000 knowledge artifacts in the network in order to reduce the computation time. As a result, the publication time of these artifacts is also uniformly distributed, measured by the number of artifacts already added to the network.

Figure 6.3 shows the three parameters $\hat{\alpha}^+, \hat{\beta}^+, \hat{\tau}$ of the model PA-RD-NCOAUT (cf. Eqs. (6.1) and (6.5)) estimated according to Eq. (6.17), along with their standard errors $\sigma(\hat{\boldsymbol{\theta}}) = \sqrt{\mathrm{Var}(\hat{\boldsymbol{\theta}})}$ calculated according to Eq. (6.14). Each of the twenty points in each of the plots corresponds to a parameter, estimated based on the addition of 250 knowledge artifacts to the network according to Eq. (6.17). The knowledge artifacts are divided into twenty batches according to their publication order. That is, each of the 250

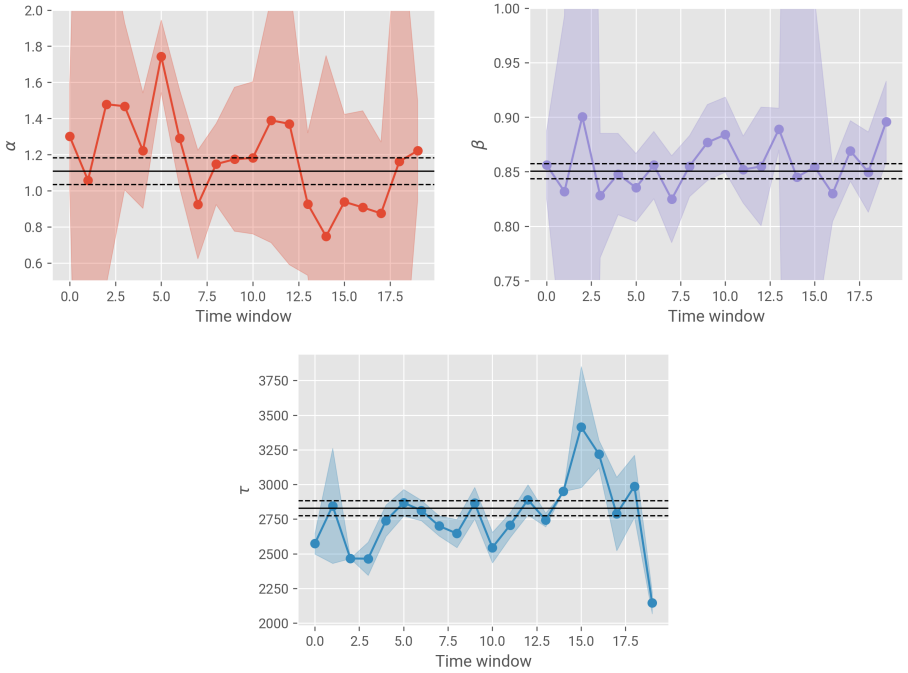


Figure 6.3: Parameters with the standard error estimated based on event samples in 20 consecutive periods in the evolution of the network. Each sample has the same number of growth events, $n = 250$. The black line corresponds to the estimate over the full growth period, with dashed lines showing the standard error. The model is **PA-RD-NCOAUT** fitted for the Nuc. Phys..

artifacts in an earlier batch is published before every artifact in a later batch. Note that although we consider a small sample of artifact additions, all the variables in the model (e.g., the number of citations $j(t)$ of cited artifact in Eq. (6.5)) are computed based on the whole network $G(t)$ at time t, where t corresponds to the time when the corresponding knowledge artifact is added to the network.

We also estimate the model parameters based on the addition of all 5000 selected knowledge artifacts. The estimate of each parameter is shown in the corresponding plot in Fig. 6.3 as a solid black line, along with the error shown as a dashed black line. For the parameter α^+, we see that the estimate and its range of the standard error based on 5000 artifacts is within the error range of all but one estimates based on the smaller sub-samples. The outcome is similar for β^+, for which only two out of twenty estimates based on a sub-sample do not cover the estimate based on all 5000 artifacts. With this

outcome, we can conclude that the maximum likelihood estimates of the parameters α^+ and β^+ and their errors are valid.

For the parameter τ, we see a problem with the error estimation. Firstly, the magnitude of the error is approximately the same for 250 observations, and for 5000 observations. Secondly, most of the error ranges based on the sub-samples do not cover the estimate based on all 5000 selected artifacts. This means that some assumptions that make the parameter estimation normally distributed are not met. Fortunately, from visual inspection of the corresponding plot in Fig. 6.3, we see that the estimate of τ based on the 5000 selected artifacts is in between the twenty estimates based on the sub-samples. This means that, even though we cannot use the estimation of the parameter error for τ, we can still use the parameter estimate itself.

As the twenty parameter estimates are based on temporally ordered growth events, we can draw conclusions about the temporal trends in the growth of the collaborative knowledge network. Remarkably, a closer look at Fig. 6.3 reveals that there are no trends in model parameters over time—all parameter estimates fluctuate around a constant mean. This is not trivial, as the time period that the twenty sub-samples represent, span the whole history of the journal from 1956 to 2017. Even though the landscape of science has changed fundamentally in this period of more than sixty years span, the model parameters are more or less the same. We believe, that the outcome would be different, if the evolution of the network was measured in real time, instead of in terms of newly added publications [149].

Growth of JHEP network Above, we have established that the fitting of coupled growth models for citations can be efficiently done based on a sample of growth events, instead of basing it on the full growth history of a collaborative knowledge network. Let us now investigate the fits of different model formulations on the example of JHEP data set. Table 6.1 shows the outcomes for eleven model formulations. As above, we perform the maximum likelihood estimation based on 5000 artifacts chosen uniformly at random. As a baseline, we take a naive model of uniform citation formation (marked `UNIF` in the table). In this model, we assume that each observed citation from a newly added artifact is drawn to an already existing artifact uniformly at random. We report the resulting log-likelihood of the model per edge, i.e., $\frac{\ln \mathcal{L}}{|E^{pc}|_s}$, where $|E^{pc}|_s = 60131$ is the number of edges in the considered sample. Given that we are considering the addition of 5000 artifacts, we see that each of these, on average, cites $|E^{pc}|_s/5000 \approx 12$ artifacts when it is added to the network.

Table 6.1: Fitting growth models to the network of JHEP based on a sample of 5000 artifacts that create $|E^{pc}|_s = 60131$ citations.

Model m	$\frac{\ln \mathcal{L}}{\|E^{pc}\|_s}$	AIC_m	w_m	$\alpha^{+/*}$	τ	$\beta^{+/*}$
UNIF	-3.86236	464495	0.00			
PA	-3.75232	451264	0.00	8.47 ± 0.188		
PA-RD	-3.60539	433595	0.00	2.08 ± 0.057	2735	
PA-NL-RD	-3.61012	434164	0.00	0.93 ± 0.005	2644	
PA-RD-NAUT	-3.60267	433270	0.00	0.77 ± 1.000	2749	0.82 ± 1.000
PA-RD-NCOAUT	-3.60015	432968	1.00	1.11 ± 0.074	2828	0.85 ± 0.007
PA-RD-MAXCOAUT	-3.60099	433068	0.00	1.17 ± 2.546	2786	0.86 ± 0.299
PA-RD-NPUB	-3.60229	433224	0.00	1.42 ± 0.437	2764	0.90 ± 0.012
PA-RD-MAXPUB	-3.60313	433325	0.00	1.57 ± 1.584	2779	0.92 ± 0.611
PA-RDxNCOAUT	-3.60167	433150	0.00	2.38 ± 0.471	2794	0.14 ± 0.010
PA-RDxNPUB	-3.60313	433326	0.00	2.27 ± 0.051	2740	0.10 ± 0.006

Next, we fit three model formulations without the coupling term, i.e., we only consider the dependence of new citations on the previous history of the citation network. The three models correspond to the three formulations of the citation component in Eqs. (6.4) to (6.6). By comparing the log-likelihoods and the AICs, we find that the two models that include relevance decay, `PA-RD` and `PA-NL-RD`, considerably outperform the simple preferential attachment model `PA`. As the preferential attachment mechanism favours older nodes, we see that the high value of the parameter $\alpha^+ = 8.47 \pm 0.188$ in `PA` (which adds a constant term to odds of being cited) tries to compensate for the relevance decay. The two models `PA-RD` and `PA-NL-RD` estimate a characteristic relevance decay time $\tau \approx 2700$. Recall that we measure the time in terms of newly added artifacts, so the value of τ shows the number of artifacts it takes to add to the network after a given publication, before its relevance is reduced by a factor $e \approx 2.72$. Note that according to the non-linear preferential attachment model `PA-NL-RD`, the preference is sub-linear on the number of previous citations, as $\alpha^* = 0.93 \pm 0.005 < 1$. Out of the two models with relevance decay, `PA-RD` fits the data better.

We now consider the additive coupling form defined in Eq. (6.1) between the citation component `PA-RD` and all presented formulations of the social component. We refer to these as `PA-RD-<soc>`, where `<soc>` is one of the formulations of the social component. All these formulations perform better than the models without coupling, as reflected in the corresponding values of the log-likelihood and the AIC. Among these, the model that best fits the data is `PA-RD-NCOAUT`, which accounts for the total number of coauthors of the authors of the cited artifact. This model is better than the other ones

that account for the team size (PA-RD-NAUT), the number of publications previously written by the authors (PA-RD-NPUB), as well as the formulations that only consider the network position of the most prominent author (PA-RD-MAXPUB and (PA-RD-MAXCOAUT)). The mixture parameter $\beta^+ = 0.85 \pm 0.007$ in the additive coupling shows that approximately 85% of the influence on the growth of citations is associated with the number of previous citations and 15% of the influence is associated with the social component, i.e., the number of previous coauthors.

Finally, we consider the multiplicative coupling form, defined in Eq. (6.3), between the citation component PA-RD and two formulations of the social component, NCOAUT and NPUB. Both models result in strongly sub-linear effect of the social component on the preferential attachment growth of the citations, with $\beta^* = 0.14 \pm 0.010 > 0$ for NCOAUT and $\beta^* = 0.10 \pm 0.006 > 0$ for NPUB. However, the effect of the social component is statistically significant.

Having eleven different models fitted to the data, we can perform model selection based on the relative likelihood w_m of the models defined on the basis of the AICs in Eq. (2.21). As a result, we find that PA-RD-NCOAUT is the sole selected model with $w_m = 1$.

Growth of twelve empirical networks Above, we have compared different models of the growth of citations in one data set. We have considered seven model formulations with coupling between layers of a collaborative knowledge network and three formulations without coupling, i.e., the ones that only consider the citation network. To have a baseline for comparison, we also included a model in which the cited articles are selected uniformly at random.

Here, we perform the same comparison for all twelve empirical collaborative networks introduced in Section 2.3. As before, the parameters are estimated based on a random sample of 5000 artifacts added to the network throughout its growth. The exception is RMP, which in total has 3006 publications, all of which are considered for model fitting. The number of citation edges $|E^{pc}|_s$ drawn from the considered sample of artifacts varies among the networks from 4318 for RMP to 60131 for the previously discussed JHEP. The outcomes for all considered models for each of the networks are presented in Appendix C. In Table 6.2, we present only the models that have a relative likelihood $w_m > 0$ for each of the networks.

We find that the selected models of the citations growth in most of the studied col-

Table 6.2: The selected growth models for twelve collaborative knowledge networks.

Model m	$\frac{\ln \mathcal{L}}{\lvert E^{pc} \rvert_s}$	w_m	$\alpha^{+/*}$	τ	$\beta^{+/*}$
PR					
PA-RD-NCOAUT	-4.00314	1.00	0.95 ± 0.247	5185	0.90 ± 0.020
PRA					
PA-NL-RD	-4.13827	1.00	1.14 ± 0.224	8411	
PRC					
PA-NL-RD	-3.92745	1.00	1.09 ± 0.007	4860	
PRE					
PA-NL-RD	-4.14851	1.00	1.21 ± 0.013	9706	
RMP					
PA-RD-NPUB	-2.84885	0.71	0.47 ± 0.086	480	0.85 ± 0.124
PA-RD-MAXPUB	-2.84909	0.26	0.45 ± 0.032	479	0.84 ± 0.016
PA-RDxNPUB	-2.84959	0.03	0.74 ± 0.315	480	0.19 ± 0.021
JHEP					
PA-RD-NCOAUT	-3.60015	1.00	1.11 ± 0.074	2828	0.85 ± 0.007
PR-HEP					
PA-RD-NPUB	-3.99631	1.00	0.65 ± 1.403	7347	0.87 ± 0.265
Phys. Lett.					
PA-RD-NPUB	-3.56096	1.00	0.40 ± 0.186	2991	0.81 ± 0.075
Nuc. Phys.					
PA-RD-NPUB	-3.61780	1.00	0.51 ± 2.072	3278	0.85 ± 0.129
PAT 320					
PA-RDxNCOAUT	-3.35807	0.93	0.77 ± 0.061	1964	0.05 ± 0.026
PA-RDxNPUB	-3.35816	0.07	0.77 ± 1.281	1958	0.05 ± 0.028
PAT 424					
PA-RDxNPUB	-3.44645	1.00	0.49 ± 0.630	2817	0.07 ± 0.675
PAT 703					
PA-RD-NAUT	-3.36720	1.00	0.46 ± 0.054	2740	0.93 ± 0.024

laborative knowledge networks are coupled models that account both for a citation component and a social component. All three exceptions are journals of physics published by the American Physical Society. These are the journals PRA, PRC and PRE. For all three, the model of non-linear preferential attachment PA-NL-RD is selected. We do not know whether coupled models would be selected for these journals as well, if we considered the citation component PA-NL-RD in coupled models. We chose not to investigate these scenarios as for many of the studied networks the exponent of the preferential attachment α^* is not significantly different from one (i.e., $\alpha^* = 1$ is within the estimated error range). Even for one of the three journals, PRA, the exponent $\alpha^* = 1.14 \pm 0.224$ is not significantly different from one, which corresponds to the linear preferential attachment PA-RD. Notably, these findings are in contrast with the claims by Golosovsky and Solomon, who argue that citation growth follows the super-linear preferential attachment, with the exponent $\alpha^* \in [1.25, 1.3]$ [80].

In most of the networks for which the selected models are coupled, the form of this coupling is the additive one. The effect of the social component in these networks, measured by its mixture weight $1 - \beta^+$, ranges between 7–19%.

The main exceptions, for which the multiplicative coupling form is selected, are PAT 320 and PAT 424. In RMP, the model PA-RDxNPUB with multiplicative coupling is selected among three models with a low likelihood weight $w_m = 0.03$. Thus, the multiplicative coupling is primarily selected in the patent networks, and not in the scientific networks. However, the parameter $\beta^* = 0.07 \pm 0.675$ of the multiplicative coupling is statistically insignificant for PAT 424, as it includes the value $\beta^* = 0$ at which the effect of coupling diminishes. For PAT 320, the parameter β^* is significant, but has a low value ($\beta^* = 0.05 \pm 0.026$ for the model PA-RDxNCOAUT with $w_m = 0.93$). The third patent network PAT 703 is a notable outlier. First, different from the other patent networks, additive coupling is selected for PAT 703. Second, it is the only studied network for which the selected social component is the team size NAUT.

Given the choice between the social component based on the whole team of authors, or on the most prominent author, the model based on the whole team is selected. Only in the case of RMP, the model PA-RD-MAXPUB that includes the highest number of publications among authors is selected with model likelihood $w_m = 0.26$, along with its counterpart for the whole team PA-RD-NPUB with $w_m = 0.71$.

To summarise, we have found that in most of the networks, the coupled growth of

citations that accounts for a social component has a higher likelihood to explain the observed citations than the simple growth model that only considers the previous history of the citation network.

6.5 Conclusion

In this chapter, we have studied the growth of the citation layer of collaborative knowledge networks. We started by reproducing the thoroughly studied models of growth based on preferential attachment that treat the citation network in isolation. In these models, the addition of new citations depends on the previous history of the citation network. Building on this, we have formulated coupled growth models for citations, such that the addition of citation edges depends on two layers of a collaborative knowledge network. Specifically, we add a social component of growth that encodes the effect that the authors of an artifact have on the likelihood of the artifact to be cited.

By fitting different model formulations, we have found that in the majority of the studied empirical networks, the best model implements the coupling with the social component. Among these, the additive coupling form is selected in most of the cases, with the mixture weight of the social component between 7% and 19%. From the selected models, we learn that the aggregated network positions of all the authors of an artifact explain the citation growth better than the network position of only the single most prominent author. This partially contradicts Merton's conjecture that the scientific community tends to attribute a publication merely to the most prominent author. A model with the data of many authors is more complex than a model that incorporates data one author, even though the difference in complexity is not easy to quantify. Hence, one may say that these two models cannot be compared without adding some suitable penalty. We leave it to the future research to investigate whether this difference in model complexity affects the presented results in any way.

Our analysis was limited exclusively to citation growth, meaning that we took the authorship relations observed in the data as a given. We did not question how the network layers, on which the social components were defined, are formed. We have outlined a model for the simultaneous growth of the whole collaborative knowledge network. Investigating this comprehensive model will allow understanding the *co-evolution* of the citation and authorship edges. Based on this, we may also be able to understand the feedback mechanisms involved in the dynamics of the networks.

Furthermore, we might gain insights into the constituents of a successful career path of a researcher. For example, we may be able to learn whether it is a highly cited publication that makes a researcher more attractive as a coauthor, thus making her more central in the coauthorship network, or whether it is the prominent coauthors that help to propel the number of citations towards her publications.

Among the limitations of our study is the fact that our models did not account for the topical community structure in the collaborative knowledge networks. It is known that artifacts in larger research areas tend to have more citations [104, 200]. We do not know how the outcomes would have changed, if we had included the effect of topical similarity in the growth of citations [30].

To summarise, there are three important contributions made by this chapter:

1 We have formulated coupled growth models of multi-layer networks, which allows studying the inter-dependent dynamics of the layers. In the context of collaborative knowledge networks, different layers—such as the citations between publications, or the coauthorship layer—have been commonly studied in separation, ignoring the influence of one layer on the dynamics of the other.

2 Contrary to the common approach of judging the growth models based on highly aggregated properties of static networks (e.g., the degree distribution), we have followed a more rigorous statistical approach that assesses the growth models based on their likelihood to observe the growth process itself. That is, we judge the models based on their likelihood given the sequence of individual edges being added to the evolving network. To our knowledge, only two previous studies by Leskovec *et al.* [109] and by Medo [124] employed a comparable microscopic approach to growth modelling.

We advance this approach to growth modelling of networks by providing standard errors for the model parameters. While error estimation is deemed necessary in most fields involving statistical inference, it has so far been largely neglected in the community studying network growth. However, the estimation of parameter errors is necessary if we want to determine the statistical significance of the outcomes.

3 We have shown that growth models can be studied in an efficient and scalable manner for large networks by means of sampling of growth events uniformly throughout the growth history of the network. Medo previously showed that

the accuracy of model estimation grows with the time window in which the growth of a network is considered [124]. The difference is that in our approach, probing the network growth throughout the growth history implicitly accounts for the growth process in between the probes, as the network properties that define the models bear the marks of this intermediate growth.

Chapter 7

Social influence on attention decay

Summary

In this chapter, we examine how attention towards knowledge artifacts—measured by means of the citation rate—is distributed over time. It has been shown that the citation rate on average follows a certain pattern: in the first period after publication, the citation rate grows up to a maximum value, after which it decays over time [149]. We study how the citation rate dynamics varies depending on the positions of knowledge artifacts' authors in the collaborative knowledge network. We find that, on average, it takes less time to achieve the highest citation rate for authors who either published more publications, or who have more coauthors. However, for these authors, the decay in the citation rate is also faster, meaning that their publications are forgotten sooner.

This chapter has been written specifically for this dissertation. Contains unpublished results by Nanumyan V., Zingg Ch., Scholtes I., and Schweitzer F. Contains results from the Master's thesis of Ch. Zingg performed under supervision of V. Nanumyan and F. Schweitzer. VN analysed and interpreted the results with the exception of fitting the exponential decay, which was provided by Ch. Zingg.

This chapter is motivated by two recent studies [149, 165]. In the first study by Sarigol *et al.* [165], scientific publications in the field of Computer Science are considered successful if five years after publication they are among the top 10% of the most cited articles published in the same year. Sarigol *et al.* find that the authors of successful publications are considerably more central in the coauthorhip network than the authors of non-successful publications. This finding holds for different centrality metrics, such as degree, k-core, betweenness. Moreover, they find that the centrality of the authors in the coauthorship network can predict the future success of their publications. In particular, a simple statistical classifier performs remarkably well. It marks as "successful" the publications by authors who are among the 10% most central authors in the coauthorship network, according to different centrality metrics (and combinations of those). The outcome of the best classifier achieves 60% precision, which is an increase of a factor of six compared to the random guess. These findings are in line with our own outcomes presented in the previous chapter. We have shown that in many citation networks, the probability for a publication to be cited grows with the number of coauthors (degree centrality in coauthorship network).

The second study by Parolo *et al.* [149] addresses the dynamics of attention towards publications. The number of citations is taken as the proxy of attention. To study how attention towards a publication changes over time, Parolo *et al.* look into the dynamics of the citation rate of the publication, i.e., the number of citations per time unit. They identify that the average dynamics of the citation rate over time has two stages. In the first stage, which lasts for 2–7 years, the rate of citations grows, reaching a peak. After the peak citation rate is reached, the second phase starts in which the citation rate decays over time. This average dynamics of the citation rate is observed in multiple scientific fields and over a long period of time. Specifically, the two-stage dynamics is shown in physics, chemistry, biology and medicine over the course of thirty years between 1960 and 1990.

In their study, Parolo *et al.* focus on long-term trends in attention decay. They find that the average time to reach the peak citation rate decreases for articles published in later years. Similarly, the decay rate increases for the articles published in later years. That is, the decay becomes faster over time, meaning that nowadays, the publications are forgotten sooner than they were in the past. Given the exponential growth of new publications every year, the authors attribute these trends to the finite capacity of scientists to follow new scientific literature. In support of this claim, they show that both the time to the peak citation rate and the characteristic decay time stay approximately the same throughout the whole considered period, if time is measured

in terms of newly published articles and not in years. Thus, older publications are forgotten not after a certain amount of time has passed, but after a certain number of new articles has been published. This is also in agreement with the outcomes of the last chapter, where we obtained stable parameter estimated for a growth model with time measured in terms of new publications.

The finding that the total number of citations a knowledge artifact receives is positively influenced by the social position of its authors indicates that there is higher attention towards publications of more central authors. If we assume that scientists have a limited attention capacity, it is plausible to attribute the increase in attention to a diffusion of information about the artifacts, which is a function of social communication between scientists [87]. That is, scientists learn about some of the publications that they later cite through communication with other scientists. Without this communication, they may not have ended up finding and citing these publications. Thus, we conjecture that at least part of the attention towards publications is due to communication between authors. We can further assume that the authors who are more central in the coauthorship network have communication channels to more scientists. Hence, the information about artifacts by more central authors can reach a higher number of scientists through social communication than the information about artifacts by less central authors.

Furthermore, we ask whether this increase in attention happens within a particular time frame: is it mostly in the early period after publication, is it spread uniformly over time, or does it happen after a rather long time has passed since publication? This question can be answered by investigating the differences in citation rate dynamics between the publications by authors who are central in the coauthorship network and publications by authors who are not central.

7.1 Citation trajectories

If $k_i^{\text{in}}(\delta)$ is the in-degree of the knowledge artifact $i \in V^p$ at time δ in the network of artifact citations G^{pc} (cf. Section 2.2), and t_i is the time of publication of the knowledge artifact i, we can write the citation rate of the artifact $t = \delta - t_i$ time units later as

$$c_i(t) = \frac{k_i^{\text{in}}(\delta + \Delta t) - k_i^{\text{in}}(\delta)}{\Delta t}. \tag{7.1}$$

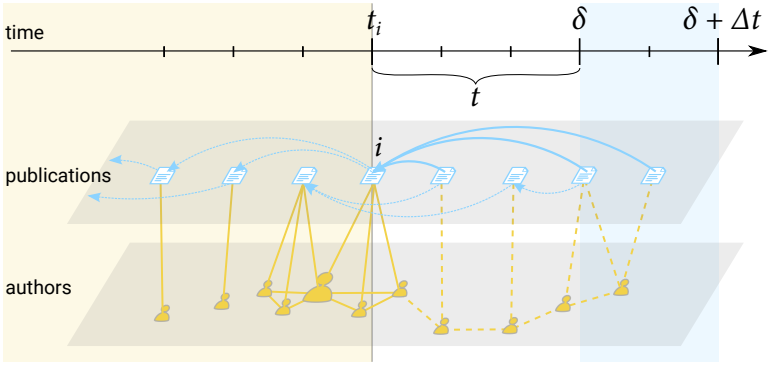


Figure 7.1: Timeline of publications with the citations to older publications and the corresponding authorship and coauthorship relations.

This is illustrated in Fig. 7.1, where the citation rate of artifact i after time t since publication is the number of citations made during the time window Δt marked by the blue shade. The time series $c_i(t)$ is called *citation history* or *citation trajectory* of artifact i [149].

We are interested in comparing the dynamics of the citation rates among knowledge artifacts. The total number of citations—and by extension, the absolute values of the citation rates—vary strongly among artifacts. Hence, we follow the example of Parolo *et al.* and we normalise the citation trajectory of each knowledge artifact by its peak value $c_i^{\max} = \max_t(c_i(t))$,

$$\tilde{c}_i(t) = c_i(t)/c_i^{\max}. \tag{7.2}$$

The left panel in Fig. 7.2 shows five exemplary normalised citation trajectories $\tilde{c}_i(t)$ of publications in PRA. We see that the patterns in citation trajectories of individual publications are quite different. The blue and gray trajectories follow a similar pattern: they reach the peak rate quite early, followed by a decrease over time. In case of the yellow trajectory, the peak is also reached rather early on and is followed by a period of low citation rate, which is in turn followed by another increase to a lower peak with subsequent decay. The red trajectory slowly grows over time, achieving a rather late peak, 12 years after publication, followed by a rapid drop in the citation rate. Lastly, the purple citation trajectory corresponds to a relatively steady citation rate over the span of almost 30 years. In the following, we will discuss the average citation trajectory in a network. With these example trajectories, we illustrate that there is high variability in the citation trajectories of individual publications, and that the average citation trajectory will not generally apply to individual publications.

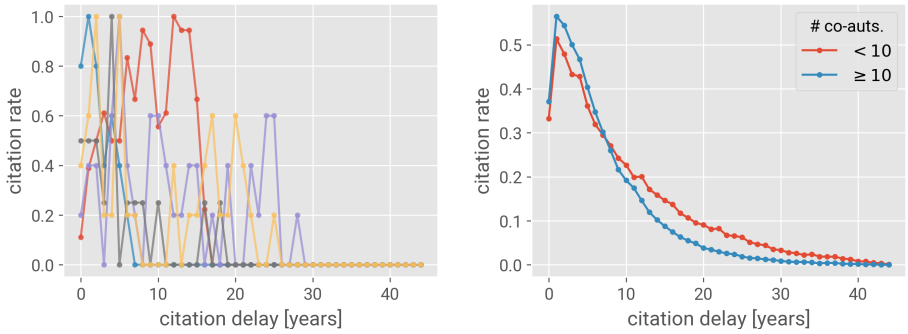


Figure 7.2: Normalised citation rate $\tilde{c}_i(t)$ of publications in PRA. (Left) $\tilde{c}_i(t)$ shown for five publications selected at random. (Right) the mean normalised citation rate $C(t)$ for two categories: publications by authors with fewer than 10 coauthors in total, and publications by authors with 10 or more coauthors prior to the considered publication.

7.1.1 Mean normalised citation trajectory

In order to get an insight about the average dynamics of the citation rates, let us calculate the mean normalised citation trajectory over all artifacts in the network as

$$C(t) = \frac{\sum_{i=1}^{N} \tilde{c}_i(t)}{N}, \tag{7.3}$$

where $N = |V^{\mathrm{P}}|$ is the number of knowledge artifacts.

We do not consider the artifacts published within the last 5 years covered in the data, because, the lengths of their citation trajectories are too short, so that the two characteristic phases in the citation trajectories cannot be observed. Dropping a longer time period could have resulted in a better representation of the mean citation trajectory, but that would be mean losing a considerable number of artifacts published in recent years. For instance, for those 5 discarded years, the number of considered publications drops from 69147 to 54782 for PRA, which means that more than 20% of articles in this journal are published during the last 5 years. Instead of choosing a longer period to omit, we discard the older artifacts for which the citation rate was still growing at the latest time step observed in the data. That is, we only consider knowledge artifacts which are already in the decay phase. This approach mostly eliminates artifacts published closer to the end of the data set. However, it also eliminates older publications that are called "sleeping beauties"—publications that remain unnoticed for a prolonged period of time, only to become frequently cited

afterwards [27, 201]. However, sleeping beauties are extremely rare, so discarding them will not affect our statistical outcomes. For instance, Van Raan identified only 0.04% of the articles published in 1988 as sleeping beauties.

In the beginning of the chapter, we have posed the question of whether the centrality of authors in the coauthorship network affects the dynamics of the citation rate of their publications. Having defined the mean normalised citation trajectory for a set of artifacts in Eq. (7.3), we can now get a first insight about this question. Let us look at the mean trajectories of the most successful publications in a journal, defined as top 10% most cited ones, divided into two groups according to the degree centrality (i.e., the number of coauthors) of the their authors. We calculate the centrality of the authors at the point when the publication is made. That is, for artifact i added at the time t_i, we build the coauthorship network $G^{aa}(t_i)$ based on the publications made earlier, as shown by the yellow shading in Fig. 7.1. We have found in Chapter 6 that the aggregated number of coauthors for all authors of a publication is a better predictor for the citations of the publications than the number of coauthors of the most central author. Hence, we characterise a publication by the total number of distinct coauthors of its authors at the time of publications as defined in Eq. (6.8) on Eq. (6.8).

Now, we put the artifacts that are characterised by fewer than the median number of coauthors in one group, and the rest in the second group. For instance, the median number of coauthors of the authors is 10 for the 10% most cited publications in PRA. The resulting mean normalised citation trajectories $C(t)$ for these two groups of artifacts in PRA is shown in the right panel of Fig. 7.2. The red trajectory, corresponding to less central authors, is averaged over 2109 artifacts and the blue trajectory for more central authors is based on 3786 artifacts. We see that the mean citation trajectory differs for the two groups of publications. For more central authors, a larger share of total citations happens in the earlier time after publication. Also, the decay phase after the peak citation rate is steeper for more central authors. In the following, we will quantify the difference in citation trajectories. We will study the effect of author centrality on the shape of citation trajectories. Specifically, we will study the time t_i^{peak} it takes to reach the highest citation rate to represent the first phase of the citation trajectory of artifact i and the characteristic decay time to represent the second phase of the trajectory. For the latter, we first need to define the functional form of the decay phase.

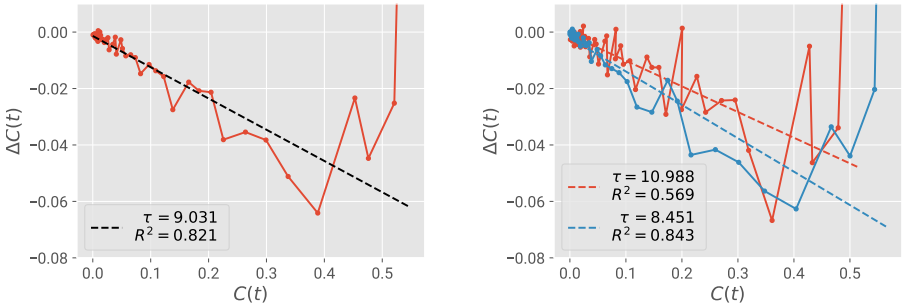


Figure 7.3: (Left) the empirical relation between the change $\Delta C(t)$ in mean citation rate and the citation rate $C(t)$ for the top 10% most cited publications in PRA and the linear fit according to Eq. (7.5). (Right) the same relation for the two groups of publications corresponding to (red) fewer than the median number of coauthors and (blue) median or greater number of coauthors.

7.1.2 Functional form of decay

Parolo *et al.* suggest two candidates for the functional form of the decay phase of the citation trajectories of knowledge artifacts [149]: exponential function

$$\tilde{c}_i(t) \propto \exp(-t/\tau), \tag{7.4}$$

and power law function $\tilde{c}_i(t) \propto t^{-\tau}$. They find that both functions provide good fits with low p-values. However, compared to each other by means of F-scores of curve fitting to trajectories of individual artifacts, they identify the exponential function as the better one for the majority of the artifacts.

Let us confirm heuristically that we too can describe the citation trajectories in our data by means of the exponential function. If the equation $\tilde{c}_i(t) \propto \exp(-t/\tau)$ holds, then the following also holds:

$$\frac{\Delta \tilde{c}_i(t)}{\Delta t} \propto -\frac{\tilde{c}_i(t)}{\tau}. \tag{7.5}$$

In other words, change in the citation rate over time is negatively proportional to the citation rate itself in the case of exponential decay form. The left panel of Fig. 7.3 shows the empirical relation between the change $\Delta C(t)$ in the mean citation rate and the mean citation rate $C(t)$, averaged over the top 10% most cited publications in PRA. It also shows the result of fitting a line to the relation, along with its slope $\tau = 9.03$. This slope identifies the characteristic decay time in years, which is the average time it takes for

the citation rate to drop $e \approx 2.72$ times. We fitted the line only to the decay phase by omitting the first two years of the citation trajectories (cf. Fig. 7.2). As in this chapter we are interested in the influence of the authors' network positions on the citation dynamics, it is not crucial to find the best functional form for the citation trajectories. Instead, we need a plausible parametrisation of the shape of citation trajectories, in order to carry out a statistical analysis for these shapes. Both the visual inspection of Fig. 7.3 and the high coefficient of determination $R^2 = 0.82$ indicate that we can use the exponential form of the decay in our analysis.

Based on this heuristic analysis of the decay phase, we can already measure the difference in the decay rate between the two groups of publications described above. The right panel of Fig. 7.3 shows the corresponding empirical relations between $\Delta C(t)$ and $C(t)$, along with their linear fits. We see that the characteristic decay time is approximately 8.5 years for publications made by more central authors, compared to eleven years for publications by less central authors. Hence, the citation rate decays faster for publications by more central authors.

7.2 Social influence on citation rate

Above we have found an indication that the position of authors in the coauthorship network prior to the publication of an artifact affects the citation rate dynamics of this artifact. Let us confirm this finding by means of statistical analysis. To this end, we perform linear regression analysis on the characteristics of the citation trajectories of individual knowledge artifacts, using the positions of their authors in the coauthorship network as explanatory variables. More precisely, we study (i) the relation between the time to the peak citations and the authors' centrality and (ii) the relation between the characteristic decay time and the authors' centrality.

We formulate the characteristics of the time series of the citation rate $\tilde{c}_i(t)$ as the dependent variable. We compute these time series for each knowledge artifact based on finite $\Delta t = 365.25$ days in Eq. (7.1). We also consider an alternative to the calendar time that measures time in terms of the number of knowledge artifacts in the growing collaborative knowledge network. For this case, we take $\Delta t = \Delta N$ nodes as the unit of time within which the citations are aggregated in Eq. (7.1), such that the resulting time series has approximately the same length as the corresponding time series built with respect to calendar time.

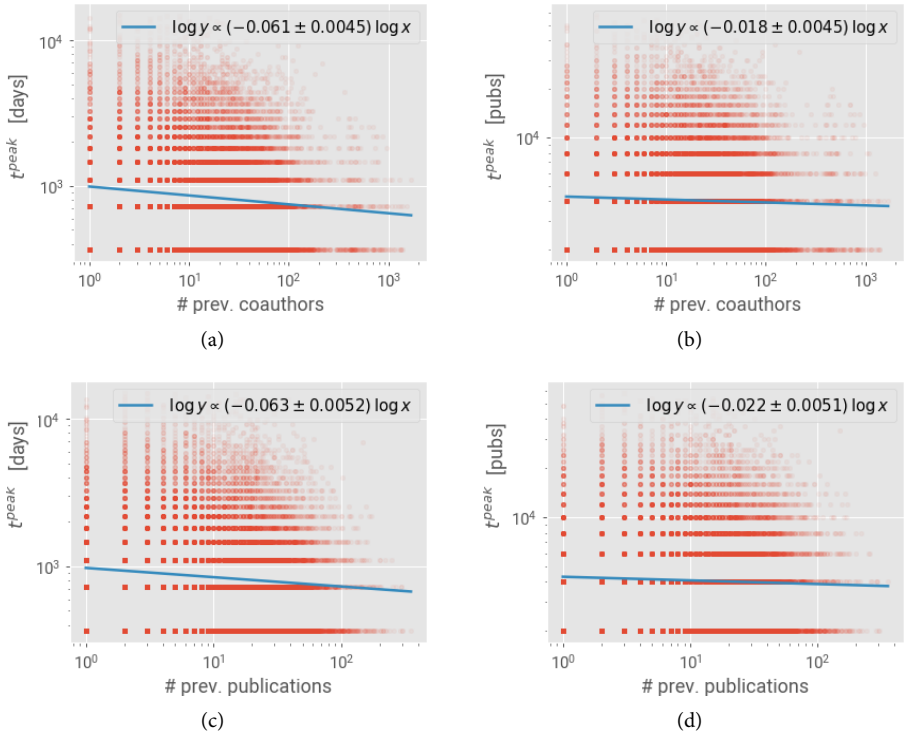


Figure 7.4: The relation between the time to the peak citation rate for an artifact and (top) the number of previous coauthors of its authors, (bottom) the number of previous artifacts by its authors. The time is measured (left) in days and (right) in terms of new publications.

In the previous chapter, we have found that in some cases, the total number of citations to an artifact can be better described by the number of other publications written by its authors, instead of the number of their coauthors. Hence, here we also perform linear regression of the citation trajectory of a knowledge artifact on the number of artifacts previously written by its authors.

7.2.1 Time to the peak citation rate

Let us first analyse the relationship between the time t_i^{peak} from the publication of an artifact i to the time of the highest citation rate, and the network positions of its authors. In order to find whether there is a significant relationship, we perform a linear analysis

for log-transformed variables:

$$\log_{10} t_i^{\text{peak}} = \beta_0^{\text{peak}} + \beta_1^{\text{peak}} \cdot \log_{10} s_i, \tag{7.6}$$

where s_i is derived from the properties of authors of the publication i. As mentioned before, we consider two variables s_i: the number of previous coauthors and the number of previous publications. As in Eq. (6.10), for an artifact i, we calculate the total number of artifacts by its authors prior to the publication time t_i as $s_i^{NP} = \left|\{\nu \mid \exists \pi_{2,i\nu}(t_i)\}\right|$, where the paths $\pi_{2,i\nu}(t_i)$ traverse the bipartite network $G^a(t_i)$ of authorship relations between artifacts and authors. Similarly, we calculate the number of distinct coauthors that the authors of i had prior to t_i as $s_i^{NC} = \left|\{\nu \mid \exists \pi_{3,j\nu}(t_i)\}\right|$ (cf. Eq. (6.8)). We choose to log-transform the dependent and explanatory variables in Eq. (7.6), as they range over several orders of magnitude.

Figure 7.4(a) shows the relation between t^{peak} and s_i^{NC} for the publications in PRA. from all the publications represented in the data, we have discarded those that were published within the last (final?) five years, as well as those for which the peak citation rate coincided with the end of the data. We find a statistically significant effect of the number of coauthors on the time to reach the peak citation rate. The log-transformed linear regression translates for the original variable as $t_i^{\text{peak}} \sim [s_i^{NC}]^{\beta_1^{\text{peak}}}$. The value of the slope β_1^{peak} has a noticeable effect on the time to the peak citation rate, given that the range of the explanatory variable s_i^{NC} covers several orders of magnitude. Specifically, the slope $\beta_1^{\text{peak}} = -0.061 \pm 0.005$ predicts that the time it takes to reach the peak citation rate for an artifact whose authors have 100 coauthors, will be on average 34% faster than for an artifact whose authors only have one coauthor. We find a similar statistical significant effect also for the number of previous publications s_i^{NP} shown in Fig. 7.4(c), with $\beta_1^{\text{peak}} = -0.063 \pm 0.005$, which means that the time it takes to reach the peak citation rate tends to be shorter if the authors of the given publication have written more publications prior to it.

So far we have measured time in terms of calendar days. As mentioned earlier, Parolo *et al.* found that there is a strong long-term trend in scientific publications, namely that the average time to the peak citation rate t^{peak} gets shorter over time [149]. Similarly, they found that the average characteristic decay time also decreases over time. This leads to a possible problem with the regression results discussed above. Specifically, the explanatory variables s^{NC} and s^{NP} that we have used in the analysis may be prone to long-term temporal trends. As a result, the dependence between t^{peak} and $s_i^{NC/NP}$

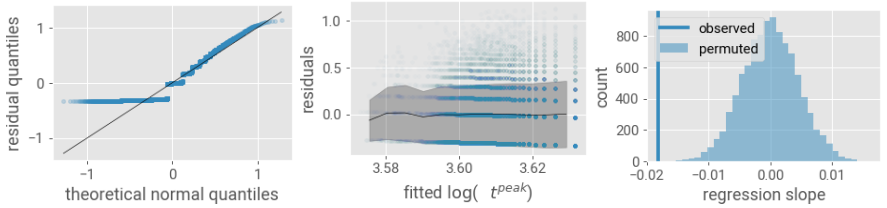


Figure 7.5: Validity of the linear regression shown in Fig. 7.4(b). (Left) quantile-quantile plot of residuals versus normal distribution, (middle) Tukey-Anscombe plot and (right) the result of permutation test based on resampling the data 10000 times.

can be a result of confounding, which threatens the interpretation that the time to the peak citation rate is affected by the network position of the authors. One way to investigate whether there is such confounding effect, is to include a term for the time of publication t_i in the regression in Eq. (7.6). Another way is to substitute the calendar time by the alternative time measured by means of knowledge artifacts added to the collaborative knowledge network. In line with the findings of Parolo *et al.* that t^{peak} and the characteristic decay time τ become approximately constant over time, as well as our own findings in Section 6.4 that the parameters of a growth model of citations are stable over time, this solves the problem of long-term trends affecting our regression results.

With this, we now repeat the regression analysis given by Eq. (7.6) by measuring the citation rate over a time unit, which is a certain number of newly published artifacts. This number is constant for each collaborative knowledge network. As mentioned above, this number is selected such that the resulting time series has the same length as the corresponding calendar-based time series. It depends on the (calendar) time span over which the network has been growing and on the number of artifacts added to the network over time.

The results of the regression analysis for t^{peak} measured in terms of new knowledge artifacts on s_i^{NC} and s_i^{NP} are shown in Figs. 7.4(b) and 7.4(d), respectively. We see that the effect is smaller than in the model where time is measured in days. For instance, for the regression on the number of previous coauthors s_i^{NC}, the slope equals $\beta_1^{\text{peak}} = -0.018 \pm 0.005$. Once we transform the logarithms Eq. (7.6) back to the original variables, we find that an artifact written by authors who collectively have 100 coauthors on average tends to reach the peak citation rate 8.7% faster than an artifact written by authors who have previously had only one coauthor.

The presented regressions have no predictive power, as determined by extremely small coefficients of determination. For instance, $R^2 = 0.001$ for the regression shown in Fig. 7.4(b) (comparative values for other cases), meaning that only 0.1% of the variance is explained by the regression model. However, this lack of predictive power does not mean that the inferred relations are not significant. While they are not useful for prediction, they show that the time to the peak citation rate is not independent of the network positions of authors. If it was, then the slopes of the regression would not be significantly different from zero. To confirm that our analysis is valid, let us look at the diagnostics for the regression residuals. As described in Section 2.4.2, for a linear regression to be valid, the residuals must be normally distributed, their variance must not depend on the explanatory variables and their expectation must be zero. To test for normality, we look at the Quantile-Quantile (QQ) plot between the observed distribution of the residuals and the theoretical normal distribution. If the observed distribution is the same as the theoretical one, the points in the QQ-plot will all fall close to the identity line. The result for the regression in Fig. 7.4(b) is shown in the left panel of Fig. 7.5. In the lower tail, i.e., for smallest negative residuals, the lowest observed quantile stretches over almost the whole negative range of the theoretical quantiles. This is due to the finite size of the time unit over which we have computed the citation rates. Non-normality of the residuals means that regression is not reliable for predictions, but it does not threaten the significance of the slope [73].

Next, we check whether the expectation of the residuals is zero and is independent from the explanatory variable. For this, we present the Tukey-Anscombe plot for the regression in Fig. 7.4(b) in the middle panel of Fig. 7.5. It shows the residuals against the predicted value of the dependent variable. The black line is the mean of the residuals for different values of the dependent variable. We see that it is close to zero for all values of the predicted dependent variable. Based on the Tukey-Anscombe plot, we can also check the condition of homoskedasticity, i.e., that the variance of the residuals are constant. To this end, we show the standard error of the residuals against the predicted value of the dependent variable. We see that it grows slightly with the dependent variable, but much less than one would assume judging from the visual inspection of the regression plot itself in Fig. 7.4(b). These plots provide only qualitative evaluation of the validity of the regression. As we are interested in the statistical significance of the dependence between the dependent and explanatory variables, we confirm the significance of the slope by means of a permutation test. That is, we reshuffle the values of the explanatory variable between different knowledge artifacts, while keeping the original values of the dependent variable. Then we perform

the regression analysis on this shuffled data. By repeating this procedure multiple times, we obtain a distribution of values of the regression analysis based on the randomised data. The right panel of Fig. 7.5 shows the outcome for the slope of the regression for 10000 randomised trials. We find that the distribution of the slope β_1^{peak} is centred around zero, meaning that there is no randomly expected effect merely from how the explanatory or dependent variables are distributed. We also find that the slope in the observed data is far outside the distribution of the randomised data, confirming the statistical significance of the identified dependence between the time to the peak citation rate and the network position of the authors. Hence, we can conclude that the network position of authors has a significant influence on the time it takes an artifact to reach highest citation rate.

7.2.2 Characteristic decay time

Let us proceed with the investigation of the characteristic decay time. As before, we perform a linear regression analysis. Here, the logarithm of the characteristic decay time τ_i of the artifact i (cf. Eq. (7.4)) is the dependent variable, and the logarithm of network position of the authors of i is the explanatory variable:

$$\log_{10} \tau_i = \beta_0^\tau + \beta_1^\tau \cdot \log_{10} s_i. \tag{7.7}$$

As above, we consider (i) the number s_i^{NC} of distinct coauthors that the authors of i had prior to the publication time t_i, and (ii) the total number of artifacts s_i^{NP} by authors of the artifact i written prior to t_i. Figure 7.6(a) shows the result of the regression on s_i^{NC} for PRA data, where the citation rates are calculated over time measured in days. For the lowest value of $s_i^{NC} = 1$, the predicted τ is approximately 10 years. For the highest values $s_i^{NC} \approx 10^3$, the predicted characteristic decay time $\tau \approx 1500$ days, or approximately four years. Note, that the predicted values are in agreement with the heuristically estimated average values shown in Fig. 7.3.

Figure 7.6(b) shows the result of the regression on s_i^{NC} for the same data, but with the time measured in new publications. In both regressions, we find that there is a statistically significant dependence between the characteristic decay time of the artifacts and the number of coauthors their authors have. The dependence is smaller when the time is measured by means of new publications. Recalling that the characteristic decay time becomes shorter over long periods of time [149], this difference in the regression

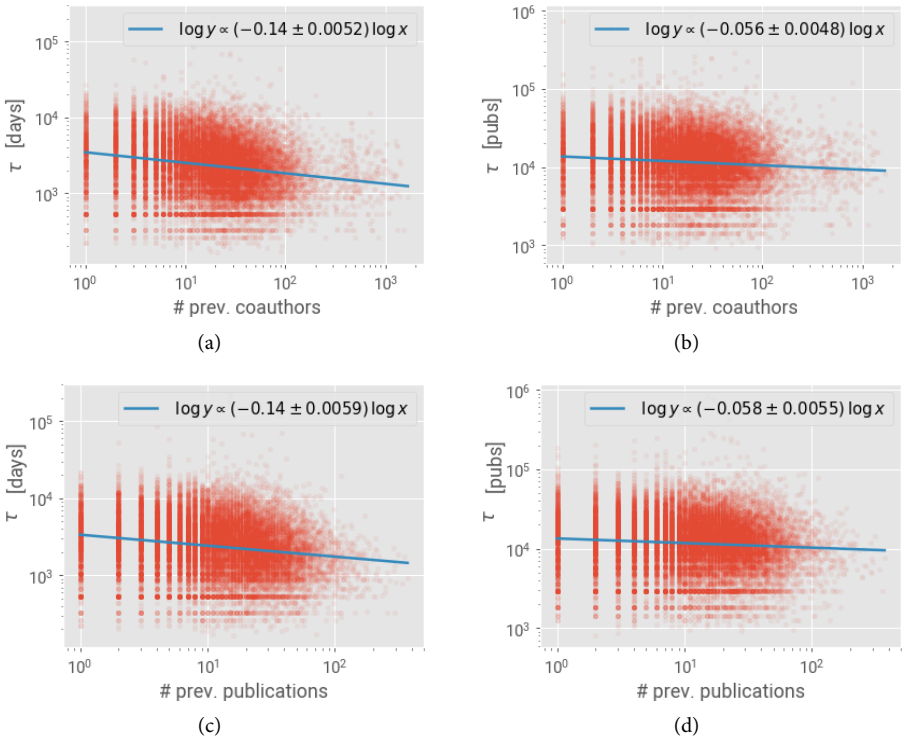


Figure 7.6: The relation between the characteristic decay time for an artifact and (top) the number of previous coauthors of its authors, (bottom) the number of previous artifacts by its authors. The time is measured (left) in days and (right) in terms of new publications.

parameters indicates that the regression suffers from confounding effects related to this trend. Performing the regression of the characteristic decay time τ_i on the number of previous publications s_i^{NP} results in similar outcomes. We find that the citation rate decays faster for the artifacts written by authors who have more prior publications, or who have more prior coauthors. Rewriting Eq. (7.7) for the original variables, we obtain $\tau_i \sim s_i^{\beta_1^\tau}$. From this, the inferred value $\beta_1^\tau = -0.056 \pm 0.005$ in Fig. 7.6(b) means that the characteristic decay time (measured by means of new publications) is on average 30% shorter for the artifacts whose authors collectively have 100 coauthors, compared to the artifacts whose authors have only one coauthor.

All four regression in Fig. 7.6 have similar residual statistics. Hence, let us now check the validity of the linear regression analysis for one example, namely the one

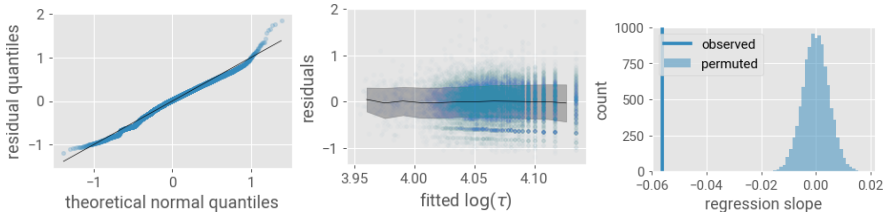


Figure 7.7: Validity of the linear regression shown in Fig. 7.6(b). (Left) quantile-quantile plot of residuals versus normal distribution, (middle) Tukey-Anscombe plot and (right) the result of permutation test based on resampling the data 10000 times.

in Fig. 7.6(b). We present the QQ-plot for the distribution of the residuals against the theoretical normal distribution in the left panel of Fig. 7.7. Overall, within one standard deviation, the observed distribution is reasonably similar to the normal distribution. Beyond that, we see that the observed positive residuals tend to be larger than predicted by the normal distribution. In contrast, the observed negative residuals below one standard deviation, tend to be smaller than predicted. Next, we inspect the Tukey-Anscombe plot shown in the middle panel of Fig. 7.7. We find that the means of the residuals are very close to zero for all values of the predicted dependent variable, with the standard deviation being almost constant. We can conclude that the conditions of the validity of the linear regression are reasonably met. To confirm the statistical significance that there is a dependence between the dependent and explanatory variables, we perform a permutation test by reshuffling the values of the explanatory variable s_i^{NC} between artifacts. Performing the regression on 10000 such randomisations, we obtain the distribution of the slope β_1^{τ} shown in the right panel of Fig. 7.7. We see that the slope inferred for the empirical data is far from the distribution from the randomised data. Hence, we can conclude that the number of previous coauthors have a significant influence on the characteristic decay time of an artifact. The more coauthors given authors have, the shorter is the characteristic decay time.

Results for twelve empirical networks So far we have only analysed the network of PRA. We repeat the regression analysis described in Sections 7.2.1 and 7.2.2 for all twelve empirical collaborative knowledge networks introduced in Section 2.3. The inferred slopes for eight regressions are presented for each network in Appendix D. Notably, when we compute the citation trajectories based on calendar time, in all but one networks, for both the time t_i^{peak} to reach the peak citation rate and the characteristic decay time τ, we find statistically significant dependence on the network position of the artifacts' authors (both on s_i^{NC} and s_i^{NP}). In all regressions with

statistically significant slopes, the dependence is negative, i.e., $\beta_1^{\text{peak}} < 0$ and $\beta_1^{\tau} < 0$. The exception is PAT 703, for which the regressions for τ result in weakly significant slopes ($0.01 < p\text{-value} < 0.05$).

When we compute the citation trajectories over time measured in terms of newly added artifacts, the slopes in the corresponding regressions become smaller. In this case, for all patent networks, the time t_i^{peak} to peak citation rate does not depend on the network positions of the authors. However, the characteristic decay time τ is still significantly affected in PAT 320 and PAT 424. For networks from the APS and INSPIRE data sets, the parameter β_1^{τ} is not significant only in JHEP and PR (for boths_i^{NC} and s_i^{NP}). The parameter β_1^{peak} is not significant only in PR-HEP, Nuc. Phys. and PRC.

7.3 Conclusion

In this chapter, we have investigated the effect of authors' network positions on the shape of the citation trajectories of their knowledge artifacts. We have found that in the majority of networks, there is a statistically significant effect. Specifically, with growing centrality of an artifact's authors, the shape of its citation trajectory tends to skew towards the time of its publication.

In the beginning of this chapter, we have conjectured that the artifacts of more central authors have more visibility due to a larger number of social communication channels of these authors. We also suggested that this increase in visibility in turn affects the citation dynamics. If this conjecture is true, then our findings indicate that this effect of increased visibility tends to happen with a short period of time after publication. One possible explanation for this is the following. When a new publication is made, the authors "advertise" it to the scientific community by presenting it in conferences and seminars, by sharing it on social media etc. This behaviour happens within a finite time period, after which the authors stop actively promoting the given publication.

This explanation is merely speculative. The analysis we have presented only shows that the network positions of authors significantly influence the citation trajectories, but it does not provide insights about the mechanisms behind this influence. In future, we may use generative modelling to learn more about these underlying mechanisms. For instance, hypotheses about these mechanisms can be formulated and tested using the framework of coupled growth models presented in Chapter 6.

Chapter 8

Conclusions

To conclude this dissertation, we summarise the main outcomes in the context of the research questions posed in Chapter 1. We also discuss the original contributions to different scientific fields and provide an outlook for future research.

In contrast with the tradition of studying citation networks and coauthorship networks in separation, we focused on questions about these networks that can only be addressed if they are considered as parts of one multi-layer network. We called this multi-layer network *collaborative knowledge network*, to reflect the fact that knowledge artifacts—patents and scientific publications—are often written collaboratively by multiple authors.

8.1 Summary in perspective

Part I of the dissertation was devoted to exploring the structural properties of collaborative knowledge networks. It is well known that these networks are characterised by pronounced community structures. We identified that new generative models are needed. Hence, we formulated the research question ***RQ 1***, which was to develop a network ensemble for collaborative knowledge networks that can account for the heterogeneities in dyadic interactions. To this end, in Chapter 3 we developed the generalised hypergeometric ensembles. These build upon two components. The first

component sets the number of possible edges between each pair of nodes in the network. A such, it incorporates combinatorial constraints on the networks. We showed how to implement a variant of the Molloy-Reed configuration model in this component. In Chapter 5, we also showed how to implement the constraints stemming from temporal ordering of events. The second component of the generalised hypergeometric ensembles, which we called edge propensities, encodes the different preferences among pairs of nodes to form edges. What we achieved is a method that can be applied far beyond collaborative knowledge networks. Any network that is a result of repeated interactions can be modelled with our ensembles. This includes co-occurrence and co-location data in human mobility, natural language processing, biology and ecology; emails; messages, likes, re-tweets and mentions in online social networks.

Having developed the generalised hypergeometric ensembles, we addressed research question ***RQ 1(a)***, which was to identify a meaningful signal in a noisy network. We followed two approaches. In the first approach, we inferred edge propensities from a network of repeated interactions by fitting the generalised hypergeometric ensembles to the observed network. The idea behind this is based on the fact that the edge propensity shows the tendency of a pair of nodes to form an edge, which goes beyond the combinatorial expectation. Then, the fitted propensities indicate the strengths of the ties between pairs of nodes, corrected for the combinatorial effects. Hence, we can regard these fitted propensities as signal, separated from the combinatorial noise. With an example of agent based model, we showed that the propensities predicted the friendship relations better than just the raw edge counts. In the second approach we followed a conventional method of backbone inference. We selected the pairs of nodes that interact more than expected from a null model. The novel contribution in this case is that the null model is not limited to combinatorial effects, but can encode preferences between nodes by means of edge propensities. This allowed us to reveal previously hidden structures in author–author citations, which we suggested may indicate special roles or anomalous behaviour among authors.

We argued that science is prone to social biases, as any other social enterprise. As the challenging goal, in ***RQ 2*** we set to explore and quantify traces of social biases in the structure of collaborative knowledge networks. For this, in Chapter 4 we developed a new statistical tool that measures how far a given observation is from the central tendency of a distribution. In Chapter 5, we exposed significant over-citations and under-citations between authors. We achieved this by modelling the citation network between authors with the generalised hypergeometric ensemble that

accounts for topical similarities between authors and respects the temporal sequence of their publications. The inferred over-citations and under-citations do not necessarily indicate social biases. However, we can safely exclude authors that cite each other as expected from the suspicion that social biases affect their behaviour. The authors who are subject of social biases are necessarily among the ones for whom we identified over-citations and under-citations. Hence, even though our procedure does not explicitly expose social biases, it makes a large step towards that.

In Part II we addressed the dynamical aspects of collaborative knowledge networks. We identified that coevolution of the different layers of collaborative knowledge networks is not studied enough, given a great amount of literature devoted to the growth of citation networks and coauthorship networks separately. Hence, we in research question ***RQ 3*** we set to define and study a model for the coupled growth of the collaborative knowledge networks. In Chapter 6, we focused on the growth of citations among knowledge artifacts in the context of the whole multi-layer collaborative knowledge network. We confirmed our expectation that such holistic models outperform the models based merely on the isolated citation network in explaining the citation formation in the majority of the studied empirical networks. In order to evaluate our coupled models, we employed the maximum likelihood estimation technique based on the formation of individual citation edges. This microscopic approach of evaluation of the growth process itself is in stark contrast with the bulk of existing studies, which evaluate the goodness of a model based on the aggregate properties of the final state of the network. We also utilised a well-known property of maximum likelihood estimation, which allows estimating errors of model parameters. While reporting statistical errors is considered necessary in most of scientific fields, it has been ubiquitously neglected in the community that studies growth of network, and in network science at large. By focusing only on citation formation, we only answered the research question partially. A more comprehensive answer would also study the formation of authorships in collaborative knowledge networks, for which we outlined the corresponding models for future research.

Based on the citation dynamics of an artifact, we addressed ***RQ 4***, which questioned how the network position of authors influences the attention towards knowledge artifacts. Specifically, we investigated the dynamics of the citation rate, which is considered as a proxy for the attention towards a knowledge artifact. Based on the existing literature, we parametrised the citation trajectory of an artifact—the time series of the citation rate—by means of (i) the time it takes for the artifact to reach the highest citation rate and (ii) the characteristic decay time after the highest citation

rate. We found a statistically significant influence of authors' network position on both parameters of the citation trajectories. However, we did not conclusively identify the extent of this influence and the mechanisms behind it.

To sum up the outcomes, throughout the dissertation we found indications that the layers of collaborative knowledge networks are strongly inter-dependent. The effects of this inter-dependence may stay unnoticed, or, in the worst case, be misinterpreted if the layers are considered in separation. Only thanks to treating collaborative knowledge networks as multi-layer networks, we were able to arrive to our conclusions.

8.2 Scientific contribution

Here we summarise the relevance of the outcomes of this dissertation in specific scientific fields. Generally, they fall under two categories, methodological and problem-oriented. The methodological outcomes mainly contribute to network science. The outcomes in the second category are all focused at quantifying and understanding the social aspects in science, thus contributing mainly to the field of scientometrics.

So far, there is one peer-reviewed publication that covers a part of the presented results. This is a position paper about inference of network backbones using generalised hypergeometric ensembles. It appeared in the proceedings of the 9th international conference on Social Informatics [37]. The generalised hypergeometric ensembles and the majority of the related results in Chapter 3 are digitally archived and publicly available [36]. Most of the content of the dissertation was presented in conferences, workshops and seminars. The generalised hypergeometric ensembles and the signed measure of deviations were presented at NetSciX, the central winter conference on Network Science. The full procedure for inferring signed relations from multi-layer networks used was presented at Complex Networks satellite meeting to Statphys26 conference and at NSF-FAST Workshop on Machine Learning. Lastly, all the methods and findings that are of relevance in scientometrics were discussed and well-received in a seminar at CWTS Centre for Science and Technology Studies in Leiden.

Complex Networks Network science deals with the issues of representing relational data as networks, as well as developing new methods for analysing these networks. In most real systems, the interactions between the elements happen in a stochastic manner. Hence, the resulting networks exhibit a mixture of regular and random patterns.

For this reason, network ensembles are an invaluable tool for studying these networks. The comparison of the network representations of real systems to network ensembles allows disentangling regular and random patterns, potentially uncovering important properties of the underlying system. The generalised hypergeometric ensembles and the related methods of backbone extraction, multiplex network regression, inference of signed deviations contribute precisely to this line of research.

Evolution of networks is another major topic in network science. Many networks representing real systems grow over time, and there is ample attention devoted to studying this growth. In Chapter 6, we contributed to multiple facets of this line of research. Firstly, we generalised a commonly used models of single-layer network growth to multi-layer networks. Secondly, we exposed the problem that statistical errors are largely neglected in this line of research. We showed how to address this issue for generative growth models by using standard methods of statistics.

Data Mining and Data Science These interdisciplinary sub-fields of Computer Science focus on methods and algorithms for finding patterns and insights in large data sets. As such, they amalgamate methods from applied statistics, machine learning and information science, as well as theories from the application domains. To deal with large data sets, a major emphasis is made on improving the computational efficiency of these methods. Topics of data mining that may benefit from our contributions include graph summarisation, collective entity resolution and anomaly detection in relational data. For instance, the latter aims at identifying patterns in data that do not comply with the expected behaviour. That is exactly what our signed measure of deviations does, whenever one's expectations are formed based on a probability distribution.

Scientometrics This field of science studies science itself by measuring scientific impact, mapping scientific fields, and developing indicators for actionable insights in policy and management of science. Management of science—for instance, allocating funds and selecting peers for evaluations—becomes more and more difficult due to increasing size and specialisation of academia.

In the recent decades, citation based indicators have been growing in popularity. They provide an easy way to rank researchers, journals and institutions. However, a debate about the adequateness of such indicators is growing as well. One of the arguments against them is that such a complex endeavour as science cannot be objectively quantified by means of highly aggregated numbers. Another important argument against

these simplistic indicators is the Goodhart's law [82]: when a measure becomes a target, it stops being a good measure. A more specific formulation of this phenomenon by Campbell exposes the problem further [33]. It states that "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor". Although there is an increasing awareness about these problems of citation based indicators, their simplicity still provides a safe haven for lazy and irresponsible decision-making. We believe that more formal studies are needed to quantify and expose the deficiencies of citation-based evaluation of science. We hope that our findings that citations are significantly influenced by social aspects will have an impact in this direction.

On a more positive note, we believe that our methods can be used in scientometrics research to quantify and better understand the inter-relations between normative and social aspects of science. Also, as we discuss in the next section, we hope to contribute to the tool-set of science managers by developing a support tool for peer selection, based on our method of identifying over-citing and under-citing researchers.

Social and Political Sciences Analysis of social networks plays a central role in quantitative Social and Political Sciences. For instance, cosponsorhsip networks formed between politicians as a result of their joint involvement in legislative initiatives are an invaluable resource for political scientists. Political polarisation a major topic of interest, which can be studied based on the topological properties of these networks. Our methodology to infer signed relations has already been used in this context in the recent doctoral dissertation of Simon Schweighofer [174].

Social scientists are as interested in negative social relations, as in positive ones. We have already discussed that there is a lack of data on signed social networks. One reason for this, we argued, is the reluctance in publicising negative relations. Another reason is that nowadays most of the data on social networks comes from online platforms, which are designed to provide a pleasant environment for interactions. Hence, they often technically limit the possibility to express negative attitudes. For example, there is the phenomenon of "liking" in online social networks, but there is often no formal way to "dislike". Our approach to circumventing this positivity bias by treating lack positive interactions as negative relations, creates an untapped potential. The abundant social networks on just positive interactions become a resource for studies on conflicts, structural balance, and other phenomena that need data on both

positive and negative social relations. Similarly, the multiplex network regression based on our ensembles provides a principled method for in-depth quantitative studies on the topic of homophily.

Systems Biology and Ecology In the recent years, a large shift in research takes place in biology. The reductionist approach of studying isolated components of biological systems is superseded by the holistic approach that studies the interrelations between these components. For instance, biological processes at a cell level involve interactions of a large number of proteins. These protein–protein interactions form a network, analysis of which is nowadays a major research topic. Uncovering the topological patterns of protein interactions is a key to understanding the micro-macro link between molecular interactions and the biological functions at the organism level. Databases of protein interactions are difficult to compile and involve complex experiments and data processing. As a result, these databases are not always reliable [189]. The experiments involve measurement errors, so statistical inference plays an important role in identifying pairs of interacting proteins. Network theoretical methods are already heavily used in this context, and we believe that our generalised hypergeometric ensembles can have an impact here as well. Researchers may be able to construct informed null models for biological network inference in a manner, similar to how we customised the combinatorial component of the ensemble for the citations between authors.

Ecologists study mutualistic networks between plants and animals in order to understand the interactions and dependencies between species in an ecological systems [11]. For instance, they use these networks to study the resilience of ecological systems. Many findings in this line depend on statistical analysis of the observed—usually error-prone—networks against statistical null models. Here as well, our framework may have an impact by providing analytically tractable models, when previously only simulations were were used.

8.3 Outlook

It is safe to say that all presented outcomes of this dissertation can be improved or extended further. In the following, we outline directions towards some of the major limitations and noteworthy extensions. Most of the original contributions are yet to be submitted to peer-reviewed scientific outlets. Hence, some of the following will be

addressed while shaping the manuscripts for submission.

One issue with the network representation of collaborative knowledge spaces, is the discrepancy between the coauthorship edges and the real underlying collaborations, as discussed in Chapter 2. We suggested that this problem is especially pronounced in the large teams, where the combinatorial exploding number of coauthorship edges cannot, in principle, represent the real collaborations between researchers. We tackled this problem by reducing the fraction of publications by very large teams of authors. In future, we would like to investigate further the relation between the network representation and the underlying real interactions in collaborative knowledge networks.

We strongly believe that the generalised hypergeometric ensembles have a largely unexplored potential. To understand this potential better, we need to thoroughly compare it with other network ensemble frameworks, such as exponential random graphs. We envision that our framework can be used in many applications where interactions are a result of combinatorial and regular mechanisms. The ensemble can be tailored to better represent the generative processes in these applications. For example, we have initial results for cosponsorhsip networks between political parties, for which we formulated the combinatorial part of the ensemble based on the bipartite structure between politicians and the political interventions. Ideally, we would also develop a method for combining different combinatorial processes, in a similar fashion to the multiplex regression for the edge propensities. Another unexplored application area is the higher order phenomena in networks. So far, only dyadic interactions are modelled by the ensembles. We would like to investigate how to incorporate group interactions, such as triadic closure, and causal paths into the framework.

With respect to the signed measure of deviations, we do not yet fully understand how the extreme values of the measure correspond to statistical significance. In particular, this must be understood for the case when the underlying distribution is discrete. Whenever we inferred significant signed relations, we chose the filtering threshold heuristically. The next step is to find an optimal threshold for a particular network. We also see the need for thorough comparison between the performance of backbone inference based on the signed measure and on a more conventional percentile-based approach. We envision an application of inferring signed relations in dynamical networks. For instance, in the context of cosponsorship-based studies of political polarisation, we can quantify the contribution of a sub-network to the polarisation. This sub-network can corresponds to individual political interventions. Hence, we can measure which interventions have a polarising and which have a depolarising

effect. For this, the signed relations can be inferred for the sub-network against the baseline provided by the whole network. Finally, we would like to carry out a detailed validation of the outcomes of inferring signed relations against sociological theories. One particular aspect of this asks for particular attention. To our knowledge, structural balance theory is only formulated for undirected signed networks. Instead, in our applications we also inferred directed signed relations. In order to assess the structural balance theory in these applications, the theory itself must be extended for the directed case.

Regarding the inference of under-citations and over-citations among authors, we constructed the ensemble based on only one measure of topical similarity. However, the multiplex network regression allows combining different similarity measures into the propensities to cite. For instance, we mentioned that we would need to include also text-based similarities in the model, in order to gain more insights about citation cartels. Hence, in future research we plan to investigate the combinations of different similarity measures with respect to their explanatory power for observed citations between authors. On top of that, the identification of certain patterns in the inferred signed networks will include network-theoretical and information-theoretical considerations. For the discussed example of identifying the interdisciplinary pioneers, one could consider the betweenness centrality of authors in the inferred network. A more sophisticated model formulation will also open the possibility for practical applications. Below, we will discuss one such application.

When modelling coupled growth, we only focused on the citation formation. In follow-up research, we would like to investigate the full model of collaborative knowledge network growth, which we merely outlined so far. This may also allow us to investigate feedback mechanisms between the normative and social aspects driving the formation of the networks.

In the last chapter, we provided the first indications that there are social influences on the citation trajectories of knowledge artifacts. We did not discuss the extent and origins of these influences. To this end, we reckon a combination of generative and discriminative modelling is needed. We used simple linear regression of parameters that characterise the shape of citation trajectories. Instead, considering the time series of individual citations towards an artifact and applying suitable methods, e.g., Cox regression, may lead to deeper insight about the social influences on the citation dynamics.

PeerSelect: a practical application One motivation for our study of over-citations and under-citations stemmed from the fact that scientific output is evaluated in a peer-review process. In order for this process to work efficiently, the recommendations by the reviewers must be based on the scientific merit of the reviewed work, and not on the social relations between the reviewer and the submitting authors. There are mechanisms in place, such as double blind reviews, that try to address this issue. However, often a reviewer is able to infer the authors of the submission, even when the names are not provided. Social biases between scientists working in the same field will also necessarily affect the peer review process, if such scientists are selected as reviewers. These biases could be potentially mitigated by a random choice of reviewers. However, this is no longer possible in many cases. The increasing specialization of research, the limited availability of experts overwhelmed by requests, the information overload to keep updated, the conflict of interests renders the number of *available* reviewers very small. Hence, the *limited* choice of reviewers makes an even larger impact on the outcome of the peer review process, which hampers its reliability [152].

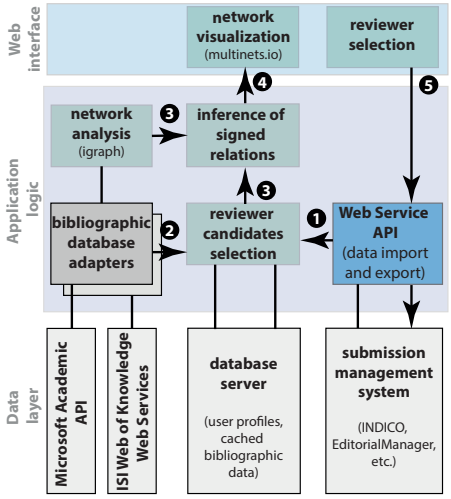


Figure 8.1: A diagram of an information system to support the peer review process in manuscript and project proposal evaluation

The role of chance in the selection of reviewers was already discussed in literature [44]. But very recently it was shown [49] that even small biases in the reviewers decisions can have big consequences in funding rates. As a consequence, such biases need to be addressed to improve both the quality of funded applications and equality among principal investigators.

To mitigate these problems, we propose a support tool for peer selection. Current systems, such as ScholarOne by *Thomson Reuters*, provide journal editors and funding agencies with a list of potential reviewers for a given submission. This list is compiled by matching research interests of reviewers and the submitting researchers, based on metadata extracted from published manuscripts. Another system is Reviewer Finder used by *Elsevier*. It matches areas of expertise and provides a very basic detection of

conflict of interests in terms of prior coauthorship. Based on text mining and keyword matching, the existing tools for peer selection ignore the social dimension that is the main source of conflict.

Our envisioned data-driven support system for the selection of reviewers is based on the framework presented in this dissertation. It is not intended to replace existing tools, but to extend them. Once a potential pool of candidate reviewers is identified by the conventional means, our tool will analyse the signed relations between these candidates and the reviewed authors. The decision makers will be able to tune the parameters of the model underlying the signed relations. In particular, they will be able to include or exclude similarity layers from the model, depending on what they consider to have a desirable or undesirable influence on the citation behaviour of the authors. For instance, they may consider correcting for the homophily due to preferred language of publications of the authors. The architecture of the proposed information system is shown in Fig. 8.1. It will rely on large bibliographic databases and web technologies. It will connect with other systems through APIs and will provide an interactive interface for the end users. Multiple publishers and funding agencies have already expressed their interest in this tool.

Appendix

Appendix A

Signed relations in Eurovision song contest

To complement Section 4.2.2, below we show results of applying the signed measure of deviations Φ defined in Eq. (4.6) to the network of votes between countries in the Eurovision song contest. Figure A.1 shows the relation between the values of Φ_{ij} and the cultural distance d_{ij} between countries i and j computed according to Eq. (4.11). While in Fig. 4.9 only the countries with at least 25 co-occurrences in the contest were considered, here all the countries with at least one co-occurrence for which the cultural dimensions are defined by Hofstede [91] are included.

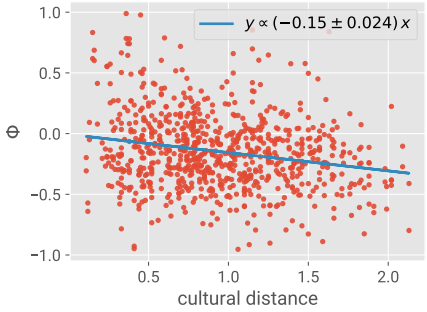


Figure A.1: The signed relations and the Hofstede's cultural distance between countries that co-participated in the Eurovision song contest at least once.

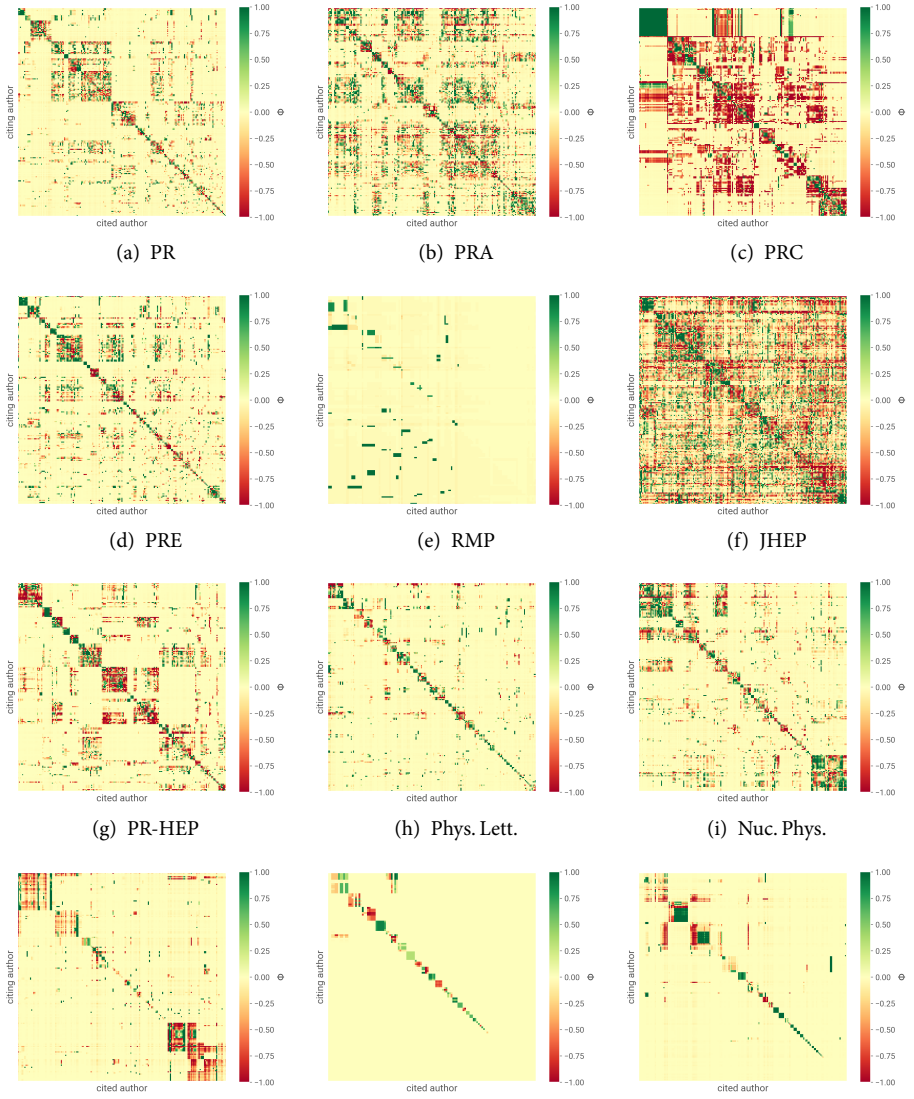
Table A.1: The most under-voting and over-voting pairs of countries in Eurovision song contest aggregated between 1975 and 2015. All the pairs co-participating at least once in the contest are considered.

Most under-represented		Φ	Most over-represented		Φ
Andorra	Georgia	-0.991	Serbia & Monte.	F.Y.R. Macedonia	1.000
San Marino	Bosnia & Herz.	-0.980	Andorra	Monaco	1.000
Andorra	Armenia	-0.972	Serbia & Monte.	Croatia	0.998
Andorra	Bosnia & Herz.	-0.961	Serbia	Bosnia & Herz.	0.993
Italy	Azerbaijan	-0.961	Turkey	Georgia	0.993
Armenia	Azerbaijan	-0.957	Moldova	Romania	0.991
Bulgaria	Italy	-0.953	Croatia	Serbia	0.989
Turkey	Serbia	-0.948	Austria	Serbia & Monte.	0.982
Ireland	Armenia	-0.947	Cyprus	Greece	0.978
Croatia	Armenia	-0.943	Slovenia	Serbia	0.977
Lithuania	Armenia	-0.943	Albania	Italy	0.977
Azerbaijan	Armenia	-0.940	Lithuania	Georgia	0.968
Switzerland	Ukraine	-0.932	Montenegro	Bosnia & Herz.	0.959
Serbia & Monte.	Turkey	-0.932	Azerbaijan	Turkey	0.954
Serbia	Turkey	-0.927	Georgia	Armenia	0.953
Czech Republic	Romania	-0.925	Georgia	Azerbaijan	0.933
Latvia	Turkey	-0.915	Turkey	Azerbaijan	0.926
San Marino	Ukraine	-0.913	Serbia & Monte.	Bosnia & Herz.	0.923
Andorra	Serbia	-0.912	Turkey	Armenia	0.910
Lithuania	Bosnia & Herz.	-0.909	Romania	Moldova	0.898

Similarly, Table A.1 corresponds to Table 4.1 with the difference that the whole network is considered and not only the countries with at least 25 co-occurrences.

Table B.1: The multiplex network regression for the citations among the 200 most cited authors on four definitions of author similarity. The networks correspond to three classes in the patent data set (cf. Section 2.3).

Similarity	$\beta_0^{(\Omega)}$	$\beta_1^{(\Omega)}$	AIC
PAT 320			
aut-in	2.486 ± 0.065 ***	1.021 ± 0.028 ***	12071.26
aut-out	1.036 ± 0.014 ***	0.714 ± 0.017 ***	17607.32
pub-in	1.972 ± 0.036 ***	0.375 ± 0.023 ***	12050.51
pub-out	1.027 ± 0.014 ***	0.309 ± 0.013 ***	17636.67
PAT 424			
aut-in	9.318 ± 236.067	1.751 ± 0.197 ***	1627.257
aut-out	1.600 ± 0.038 ***	0.475 ± 0.102 ***	2873.178
pub-in	2.504 ± 0.094 ***	1.508 ± 0.243 ***	1854.111
pub-out	1.641 ± 0.038 ***	0.665 ± 0.092 ***	2835.283
PAT 703			
aut-in	2.167 ± 0.047 ***	1.073 ± 0.052 ***	10241.04
aut-out	1.490 ± 0.022 ***	0.295 ± 0.027 ***	12051.36
pub-in	1.691 ± 0.026 ***	0.554 ± 0.033 ***	10883.58
pub-out	1.466 ± 0.022 ***	0.222 ± 0.021 ***	12137.28



(a) PR (b) PRA (c) PRC

(d) PRE (e) RMP (f) JHEP

(g) PR-HEP (h) Phys. Lett. (i) Nuc. Phys.

(j) PAT 320 (k) PAT 424 (l) PAT 703

Figure B.1: Signed relations $\mathbf{\Phi}$ based on the generalised hypergeometric ensemble with causality preserving possibility matrix $\mathbf{\Xi}$ and edge propensities $\mathbf{\Omega}$ inferred from the best fittin similarity definition. The order of rows and columns corresponds to a hierarchical clustering performed on $\mathbf{\Omega}$.

Appendix B

Signed relations among top cited authors

Table B.2: The multiplex network regression for the citations among the 200 most cited authors on four definitions of author similarity. The networks correspond to five journals in the APS data set (cf. Section 2.3).

Similarity	$\beta_0^{(\Omega)}$	$\beta_1^{(\Omega)}$	AIC
PR			
aut-in	9.534 ± 65.436	2.210 ± 0.019 ***	36044.65
aut-out	9.592 ± 65.107	2.170 ± 0.018 ***	37156.95
pub-in	3.013 ± 0.025 ***	1.135 ± 0.009 ***	27672.36
pub-out	3.253 ± 0.032 ***	1.197 ± 0.009 ***	30300.33
PRA			
aut-in	8.922 ± 49.686	2.673 ± 0.015 ***	93994.79
aut-out	2.452 ± 0.029 ***	2.183 ± 0.013 ***	103814.7
pub-in	3.123 ± 0.022 ***	1.172 ± 0.005 ***	56759.82
pub-out	2.914 ± 0.018 ***	1.238 ± 0.005 ***	69131.14
PRC			
aut-in	8.109 ± 28.018	1.789 ± 0.008 ***	134506.6
aut-out	8.462 ± 31.001	1.385 ± 0.007 ***	144462.7
pub-in	2.794 ± 0.026 ***	0.632 ± 0.004 ***	116824.6
pub-out	3.026 ± 0.036 ***	0.648 ± 0.004 ***	125867.6
PRE			
aut-in	9.560 ± 63.985	2.602 ± 0.018 ***	47552.62
aut-out	5.238 ± 0.434 ***	2.233 ± 0.015 ***	51077.64
pub-in	3.145 ± 0.025 ***	1.190 ± 0.008 ***	32414.45
pub-out	3.072 ± 0.023 ***	1.219 ± 0.008 ***	38349.14
RMP			
aut-in	0.610 ± 0.049 ***	-0.580 ± 0.164 ***	3540.834
aut-out	0.782 ± 0.058 ***	0.137 ± 0.099	3489.553
pub-in	0.659 ± 0.049 ***	-0.214 ± 0.210	3544.965
pub-out	0.759 ± 0.061 ***	0.106 ± 0.130	3490.815

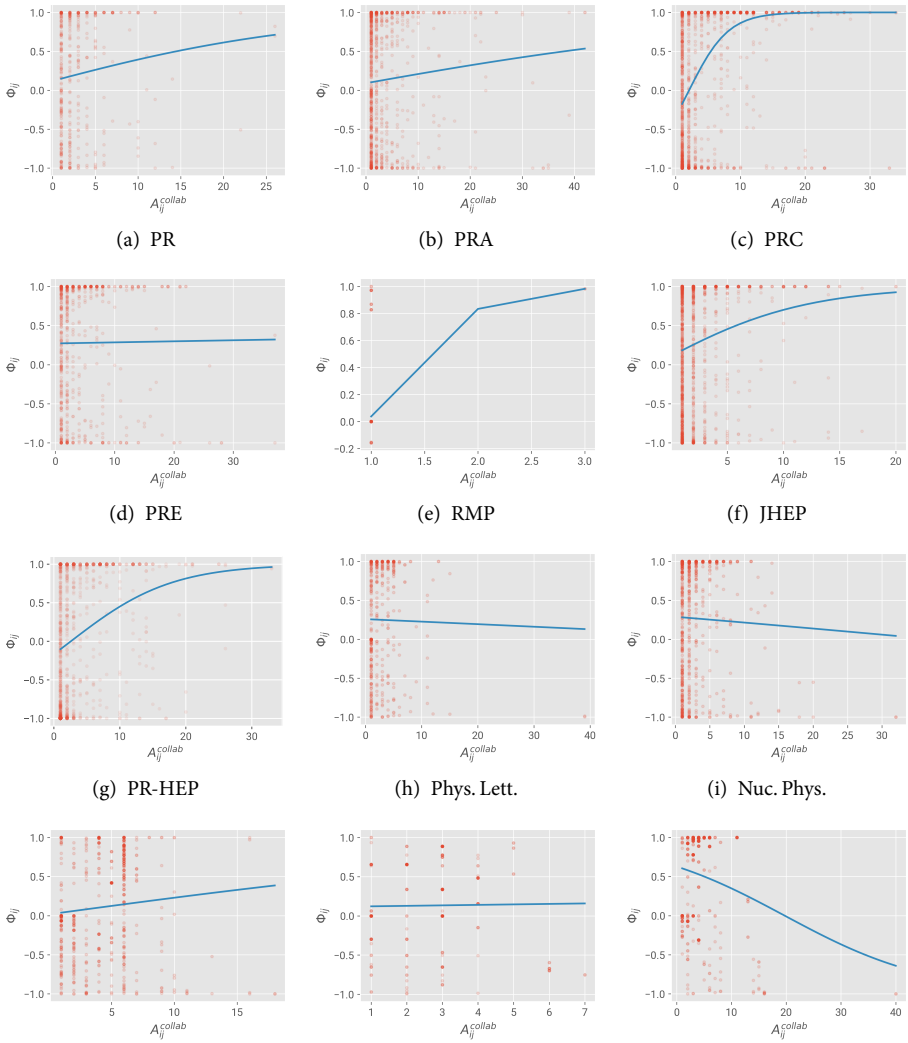


Figure B.2: Logistic regression for the sign of Φ_{ij} on the number of collaborations between authors i and j (cf. Eq. (5.11)) for the 200 most cited authors in twelve empirical networks.

Appendix C

MLE of coupled growth models

In this appendix we present the results of maximum likelihood estimation of coupled growth models that are discussed in 6. To reiterate, the growth of citation networks is studied under the assumption that the arrival of new citations to a knowledge artifact generally depend on the number of existing citations, on the age of the artifact and on a measures of social position of the contributors to the artifact. Below are the tables with the outcomes for twelve networks that are categorised according to the dataset from which the network is build. The three datasets are the Inspire-HEP, APS and the patents with disambiguated inventors [114].

C.1 APS journals

Table C.1: MLE of growth in the network of PR with $|E^{pc}|_s = 26291$.

Model m	$\frac{\ln \mathcal{L}}{\|E^{pc}\|_s}$	AIC_m	w_m	$\alpha^{+/*}$	τ	$\beta^{+/*}$
PA	-4.27084	224571	0.00	4.27 ± 0.140		
PA-RD	-4.00829	210768	0.00	1.24 ± 0.032	5133	
PA-NL-RD	-4.00878	210794	0.00	1.03 ± 1.000	4924	
PA-RD-NAUT	-4.00656	210679	0.00	0.81 ± 0.121	5145	0.88 ± 0.436
PA-RD-NCOAUT	-4.00314	210499	1.00	0.95 ± 0.247	5185	0.90 ± 0.020
PA-RD-MAXCOAUT	-4.00336	210511	0.00	0.95 ± 0.165	5208	0.90 ± 0.026
PA-RD-NPUB	-4.00431	210561	0.00	0.95 ± 6.055	5166	0.90 ± 0.399
PA-RD-MAXPUB	-4.00487	210590	0.00	0.95 ± 1.487	5171	0.91 ± 0.274
PA-RDxNCOAUT	-4.00579	210638	0.00	1.31 ± 1.637	5168	0.11 ± 0.113
PA-RDxNPUB	-4.00528	210611	0.00	1.36 ± 0.418	5202	0.12 ± 0.582

Table C.2: MLE of growth in the network of PRA with $|E^{pc}|_s = 27335$.

Model m	$\frac{\ln \mathcal{L}}{\|E^{pc}\|_s}$	AIC_m	w_m	$\alpha^{+/*}$	τ	$\beta^{+/*}$
PA	-4.38439	239696	0.00	1.44 ± 0.052		
PA-RD	-4.14304	226504	0.00	0.80 ± 0.031	8879	
PA-NL-RD	-4.13827	226243	1.00	1.14 ± 0.224	8411	
PA-RD-NAUT	-4.14224	226462	0.00	0.65 ± 0.239	8659	0.95 ± 0.062
PA-RD-NCOAUT	-4.14248	226476	0.00	0.69 ± 0.983	8574	0.97 ± 0.507
PA-RD-MAXCOAUT	-4.14234	226467	0.00	0.75 ± 0.332	9020	0.98 ± 0.012
PA-RD-NPUB	-4.14046	226365	0.00	0.62 ± 1.229	9064	0.94 ± 0.082
PA-RD-MAXPUB	-4.14057	226371	0.00	0.62 ± 0.047	9044	0.93 ± 0.009
PA-RDxNCOAUT	-4.14210	226454	0.00	0.84 ± 0.594	8808	0.05 ± 0.071
PA-RDxNPUB	-4.14138	226415	0.00	0.87 ± 0.091	8988	0.07 ± 0.009

Table C.3: MLE of growth in the network of PRC with $|E^{pc}|_s = 32358$.

Model m	$\frac{\ln \mathcal{L}}{\|E^{pc}\|_s}$	AIC_m	w_m	$\alpha^{+/*}$	τ	$\beta^{+/*}$
PA	-4.16538	269569	0.00	2.05 ± 0.072		
PA-RD	-3.92961	254313	0.00	1.10 ± 0.064	5030	
PA-NL-RD	-3.92745	254173	1.00	1.09 ± 0.007	4860	
PA-RD-NAUT	-3.92949	254307	0.00	1.04 ± 1.499	4998	0.99 ± 0.345
PA-RD-NCOAUT	-3.92936	254298	0.00	1.04 ± 0.139	4982	0.99 ± 0.063
PA-RD-MAXCOAUT	-3.92947	254305	0.00	1.05 ± 0.284	4998	0.99 ± 0.025
PA-RD-NPUB	-3.92930	254295	0.00	1.02 ± 0.038	5017	0.98 ± 0.005
PA-RD-MAXPUB	-3.92948	254306	0.00	1.06 ± 0.274	5011	0.99 ± 0.086
PA-RDxNCOAUT	-3.92946	254305	0.00	1.11 ± 0.036	4980	0.01 ± 0.007
PA-RDxNPUB	-3.92943	254303	0.00	1.10 ± 0.919	4980	0.02 ± 0.190

Table C.4: MLE of growth in the network of PRE $|E^{pc}|_s = 15355$.

Model m	$\frac{\ln \mathcal{L}}{\|E^{pc}\|_s}$	AIC_m	w_m	$\alpha^{+/*}$	τ	$\beta^{+/*}$
PA	-4.27311	131229	0.00	0.80 ± 0.036		
PA-RD	-4.15347	127557	0.00	0.60 ± 0.090	10465	
PA-NL-RD	-4.14851	127405	1.00	1.21 ± 0.013	9706	
PA-RD-NAUT	-4.15287	127541	0.00	0.47 ± 0.049	11047	0.94 ± 0.016
PA-RD-NCOAUT	-4.39126	134862	0.00	181.22 ± 1.759	13732	0.97 ± 0.016
PA-RD-MAXCOAUT	-4.38981	134817	0.00	153.98 ± 2.027	13349	0.98 ± 0.042
PA-RD-NPUB	-4.15212	127518	0.00	0.52 ± 0.250	10776	0.96 ± 0.025
PA-RD-MAXPUB	-4.15209	127517	0.00	0.52 ± 0.230	10820	0.96 ± 0.065
PA-RDxNCOAUT	-4.15264	127533	0.00	0.63 ± 0.060	10675	0.04 ± 0.074
PA-RDxNPUB	-4.37193	134268	0.00	163.60 ± 1.358	14659	0.23 ± 0.010

Table C.5: MLE of growth in the network of RMP $|E^{pc}|_s = 4318$.

Model m	$\frac{\ln \mathcal{L}}{\|E^{pc}\|_s}$	AIC_m	w_m	$\alpha^{+/*}$	τ	$\beta^{+/*}$
PA	-3.09255	26709	0.00	0.75 ± 0.051		
PA-RD	-2.85988	24702	0.00	0.62 ± 1.000	458	
PA-NL-RD	-2.85675	24675	0.00	1.27 ± 0.037	446	
PA-RD-NAUT	-2.85799	24688	0.00	0.50 ± 1.349	463	0.91 ± 0.294
PA-RD-NCOAUT	-2.85437	24656	0.00	0.53 ± 0.043	475	0.92 ± 0.015
PA-RD-MAXCOAUT	-2.85466	24659	0.00	0.52 ± 0.041	475	0.92 ± 0.014
PA-RD-NPUB	-2.84885	24609	0.71	0.47 ± 0.086	480	0.85 ± 0.124
PA-RD-MAXPUB	-2.84909	24611	0.26	0.45 ± 0.032	479	0.84 ± 0.016
PA-RDxNCOAUT	-2.85453	24658	0.00	0.68 ± 0.064	483	0.12 ± 0.020
PA-RDxNPUB	-2.84959	24615	0.03	0.74 ± 0.315	480	0.19 ± 0.021

C.2 INSPIRE journals

Table C.6: MLE of growth in the network of PR-HEP with $|E^{pc}|_s = 23089$.

Model m	$\frac{\ln \mathcal{L}}{\|E^{pc}\|_s}$	AIC_m	w_m	$\alpha^{+/*}$	τ	$\beta^{+/*}$
UNIF	-4.37538	202046	0.00			
PA	-4.18728	193362	0.00	2.14 ± 0.069		
PA-RD	-4.00319	184863	0.00	0.99 ± 0.032	7041	
PA-NL-RD	-4.00023	184727	0.00	1.10 ± 0.009	6691	
PA-RD-NAUT	-4.00249	184833	0.00	0.76 ± 0.299	7105	0.92 ± 0.042
PA-RD-NCOAUT	-3.99780	184616	0.00	0.69 ± 1.128	7461	0.89 ± 0.063
PA-RD-MAXCOAUT	-3.99813	184631	0.00	0.68 ± 0.295	7396	0.89 ± 0.029
PA-RD-NPUB	-3.99631	184548	1.00	0.65 ± 1.403	7347	0.87 ± 0.265
PA-RD-MAXPUB	-3.99745	184600	0.00	0.66 ± 0.036	7277	0.88 ± 0.009
PA-RDxNCOAUT	-4.00153	184789	0.00	1.04 ± 0.098	7265	0.08 ± 0.011
PA-RDxNPUB	-4.00128	184777	0.00	1.07 ± 0.034	7327	0.08 ± 0.009

Table C.7: MLE of growth in the network of Phys. Lett. with $|E^{pc}|_s = 11713$.

Model m	$\frac{\ln \mathcal{L}}{\|E^{pc}\|_s}$	AIC_m	w_m	$\alpha^{+/*}$	τ	$\beta^{+/*}$
UNIF	-4.04097	94664	0.00			
PA	-3.81230	89309	0.00	1.35 ± 0.055		
PA-RD	-3.57415	83732	0.00	0.69 ± 0.021	2817	
PA-NL-RD	-3.56929	83618	0.00	1.16 ± 0.215	2729	
PA-RD-NAUT	-3.57389	83728	0.00	0.61 ± 0.226	2823	0.96 ± 0.210
PA-RD-NCOAUT	-3.56816	83594	0.00	0.50 ± 0.087	2904	0.88 ± 0.972
PA-RD-MAXCOAUT	-3.56823	83595	0.00	0.47 ± 0.968	2916	0.87 ± 0.279
PA-RD-NPUB	-3.56096	83425	1.00	0.40 ± 0.186	2991	0.81 ± 0.075
PA-RD-MAXPUB	-3.56206	83451	0.00	0.39 ± 0.093	2994	0.81 ± 0.036
PA-RDxNCOAUT	-3.57159	83674	0.00	0.74 ± 0.508	2895	0.10 ± 0.054
PA-RDxNPUB	-3.56861	83604	0.00	0.79 ± 0.054	2968	0.15 ± 0.034

Table C.8: MLE of growth in the network of Nuc. Phys. with $|E^{pc}|_s = 25238$.

Model m	$\frac{\ln \mathcal{L}}{\|E^{pc}\|_s}$	AIC_m	w_m	α	τ	β
UNIF	-4.07215	205546	0.00			
PA	-3.85150	194410	0.00	2.15 ± 0.066		
PA-RD	-3.62676	183069	0.00	0.92 ± 0.042	3110	
PA-NL-RD	-3.62602	183031	0.00	1.05 ± 0.006	3067	
PA-RD-NAUT	-3.62673	183069	0.00	0.87 ± 1.563	3106	0.99 ± 0.126
PA-RD-NCOAUT	-3.62370	182916	0.00	0.67 ± 0.073	3180	0.92 ± 0.023
PA-RD-MAXCOAUT	-3.62419	182941	0.00	0.68 ± 0.474	3161	0.93 ± 0.144
PA-RD-NPUB	-3.61780	182618	1.00	0.51 ± 2.072	3278	0.85 ± 0.129
PA-RD-MAXPUB	-3.61893	182675	0.00	0.52 ± 0.037	3262	0.86 ± 0.019
PA-RDxNCOAUT	-3.62441	182952	0.00	0.97 ± 0.033	3234	0.10 ± 0.011
PA-RDxNPUB	-3.62102	182781	0.00	1.05 ± 0.212	3359	0.16 ± 0.013

C.3 Patents

Table C.9: MLE of growth in the network of PAT 320 $|E^{pc}|_s = 27425$.

Model m	$\frac{\ln \mathcal{L}}{\|E^{pc}\|_s}$	AIC_m	w_m	α	τ	β
PA	-3.41308	187209	0.00	2.75 ± 0.086		
PA-RD	-3.35871	184229	0.00	0.76 ± 0.026	1895	
PA-NL-RD	-3.35907	184249	0.00	1.05 ± 0.039	1931	
PA-RD-NAUT	-3.35847	184218	0.00	0.65 ± 0.544	1897	0.97 ± 1.579
PA-RD-NCOAUT	-3.35867	184229	0.00	0.74 ± 0.900	1887	1.00 ± 0.012
PA-RD-MAXCOAUT	-3.35868	184229	0.00	0.75 ± 0.372	1902	1.00 ± 0.005
PA-RD-NPUB	-3.35871	184231	0.00	0.76 ± 0.392	1903	1.00 ± 0.020
PA-RD-MAXPUB	-3.35871	184231	0.00	0.75 ± 0.186	1890	1.00 ± 0.015
PA-RDxNCOAUT	-3.35807	184196	0.93	0.77 ± 0.061	1964	0.05 ± 0.026
PA-RDxNPUB	-3.35816	184201	0.07	0.77 ± 1.281	1958	0.05 ± 0.028

Table C.10: MLE of growth in the network of PAT 424 $|E^{pc}|_s = 9163$.

Model m	$\frac{\ln \mathcal{L}}{\|E^{pc}\|_s}$	AIC_m	w_m	α	τ	β
PA	-3.49190	63995	0.00	1.08 ± 0.052		
PA-RD	-3.44782	63189	0.00	0.49 ± 0.058	2684	
PA-NL-RD	-3.45524	63325	0.00	1.16 ± 0.018	3128	
PA-RD-NAUT	-3.44782	63191	0.00	0.49 ± 1.000	2684	1.00 ± 1.000
PA-RD-NCOAUT	-3.44782	63191	0.00	0.48 ± 1.253	2677	1.00 ± 1.000
PA-RD-MAXCOAUT	-3.44782	63191	0.00	0.48 ± 2.727	2672	1.00 ± 1.000
PA-RD-NPUB	-3.44782	63191	0.00	0.49 ± 0.657	2682	1.00 ± 1.000
PA-RD-MAXPUB	-3.44782	63191	0.00	0.48 ± 1.000	2671	1.00 ± 1.000
PA-RDxNCOAUT	-3.44729	63181	0.00	0.49 ± 0.024	2767	0.04 ± 0.036
PA-RDxNPUB	-3.44645	63166	1.00	0.49 ± 0.630	2817	0.07 ± 0.675

Table C.11: MLE of growth in the network of PAT 703 with $|E^{pc}|_s = 11156$.

Model m	$\frac{\ln \mathcal{L}}{\|E^{pc}\|_s}$	AIC_m	w_m	$\alpha^{+/*}$	τ	$\beta^{+/*}$
PA	-3.40168	75900	0.00	1.69 ± 0.076		
PA-RD	-3.36803	75151	0.00	0.61 ± 0.374	2824	
PA-NL-RD	-3.37129	75224	0.00	0.96 ± 0.013	3732	
PA-RD-NAUT	-3.36720	75135	1.00	0.46 ± 0.054	2740	0.93 ± 0.024
PA-RD-NCOAUT	-3.36804	75154	0.00	0.59 ± 0.258	2754	1.00 ± 0.824
PA-RD-MAXCOAUT	-3.36803	75154	0.00	0.60 ± 2.062	2783	1.00 ± 1.000
PA-RD-NPUB	-3.36803	75154	0.00	0.60 ± 2.104	2772	1.00 ± 1.000
PA-RD-MAXPUB	-3.36803	75153	0.00	0.61 ± 1.000	2807	1.00 ± 1.000
PA-RDxNCOAUT	-3.36797	75152	0.00	0.60 ± 0.190	2819	0.01 ± 0.318
PA-RDxNPUB	-3.36802	75153	0.00	0.61 ± 0.030	2835	0.01 ± 0.013

Appendix D

Outcomes of regressions in Chapter 7

In this appendix, we present the results of linear regression analysis for twelve empirical collaborative networks described in Section 2.3. We report the inferred parameters β_1^{peak} and β_1^{τ} of Eqs. (7.6) and (7.7). We regress the time t_i^{peak} it takes the artifact i to reach its peak citation rate, on the total number s_i^{NC} of distinct coauthors that its authors collectively had prior to artifact i and the total number s_i^{NP} of distinct artifacts that its authors wrote prior to i. The characteristics β_1^{peak} and β_1^{τ} of the citation rate time series are computed based on calendar time units (days) and based on the alternative time measured by means of new artifacts (pubs) added to the collaborative knowledge network.

Table D.1: Results of regression analysis for three patent classes.

s_i	Time	β_1^{peak}	β_1^{τ}
PAT 320			
NC	days	-0.115 ± 0.014 ***	-0.114 ± 0.016 ***
	pubs	0.006 ± 0.013	-0.051 ± 0.015 ***
NP	days	-0.119 ± 0.013 ***	-0.121 ± 0.015 ***
	pubs	-0.003 ± 0.012	-0.057 ± 0.013 ***
PAT 424			
NC	days	-0.107 ± 0.020 ***	-0.100 ± 0.030 ***
	pubs	-0.017 ± 0.019	-0.145 ± 0.031 ***
NP	days	-0.100 ± 0.017 ***	-0.086 ± 0.026 **
	pubs	0.002 ± 0.016	-0.093 ± 0.027 ***
PAT 703			
NC	days	-0.076 ± 0.017 ***	-0.047 ± 0.023 *
	pubs	0.024 ± 0.018	-0.005 ± 0.023
NP	days	-0.075 ± 0.015 ***	-0.048 ± 0.020 *
	pubs	0.032 ± 0.016 *	0.008 ± 0.020

Table D.2: Results of regression analysis for four largest journals in Inspire-HEP data set.

s_i	Time	β_1^{peak}	β_1^{τ}
JHEP			
NC	days	-0.048 ± 0.012 ***	-0.061 ± 0.009 ***
	pubs	-0.048 ± 0.011 ***	-0.002 ± 0.008
NP	days	-0.027 ± 0.010 **	-0.026 ± 0.008 ***
	pubs	-0.034 ± 0.010 ***	0.011 ± 0.007
PR-HEP			
NC	days	-0.112 ± 0.007 ***	-0.159 ± 0.007 ***
	pubs	-0.007 ± 0.006	-0.054 ± 0.008 ***
NP	days	-0.103 ± 0.006 ***	-0.135 ± 0.007 ***
	pubs	-0.013 ± 0.006 *	-0.039 ± 0.007 ***
Phys. Lett.			
NC	days	-0.065 ± 0.012 ***	-0.080 ± 0.016 ***
	pubs	-0.084 ± 0.012 ***	-0.123 ± 0.016 ***
NP	days	-0.069 ± 0.011 ***	-0.070 ± 0.014 ***
	pubs	-0.083 ± 0.012 ***	-0.102 ± 0.014 ***
Nuc. Phys.			
NC	days	-0.062 ± 0.009 ***	-0.117 ± 0.010 ***
	pubs	-0.008 ± 0.009	-0.116 ± 0.010 ***
NP	days	-0.069 ± 0.008 ***	-0.119 ± 0.008 ***
	pubs	-0.000 ± 0.008	-0.111 ± 0.009 ***

Table D.3: Results of regression analysis for five journals published by APS.

s_i	Time	β_1^{peak}	β_1^{τ}
PR			
NC	days	-0.069 ± 0.008 ***	-0.082 ± 0.008 ***
	pubs	-0.040 ± 0.007 ***	-0.013 ± 0.008
NP	days	-0.017 ± 0.008 *	-0.032 ± 0.008 ***
	pubs	-0.023 ± 0.007 ***	-0.020 ± 0.008 *
PRA			
NC	days	-0.061 ± 0.005 ***	-0.138 ± 0.005 ***
	pubs	-0.018 ± 0.005 ***	-0.056 ± 0.005 ***
NP	days	-0.063 ± 0.005 ***	-0.142 ± 0.006 ***
	pubs	-0.022 ± 0.005 ***	-0.058 ± 0.005 ***
PRC			
NC	days	-0.019 ± 0.006 **	-0.071 ± 0.007 ***
	pubs	-0.007 ± 0.006	-0.052 ± 0.007 ***
NP	days	-0.030 ± 0.007 ***	-0.083 ± 0.009 ***
	pubs	-0.021 ± 0.007 **	-0.063 ± 0.009 ***
PRE			
NC	days	-0.023 ± 0.006 ***	-0.041 ± 0.007 ***
	pubs	-0.014 ± 0.006 *	-0.049 ± 0.008 ***
NP	days	-0.028 ± 0.006 ***	-0.036 ± 0.008 ***
	pubs	-0.019 ± 0.006 **	-0.038 ± 0.009 ***
RMP			
NC	days	-0.106 ± 0.032 **	-0.272 ± 0.054 ***
	pubs	-0.126 ± 0.033 ***	-0.337 ± 0.051 ***
NP	days	-0.099 ± 0.036 **	-0.247 ± 0.070 ***
	pubs	-0.110 ± 0.036 **	-0.281 ± 0.062 ***

List of Figures

List of Tables

Bibliography

[1] (1925). A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S. *Philosophical Transactions of the Royal Society of London B: Biological Sciences* **213(402-410)**, 21–87. ISSN 0264-3960.

[2] Abbasi, A.; Hossain, L.; Leydesdorff, L. (2012). Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. *Journal of Informetrics* **6(3)**, 403–412.

[3] Ahlgren, P.; Jarneving, B. (2008). Bibliographic coupling, common abstract stems and clustering: A comparison of two document-document similarity approaches in the context of science mapping. *Scientometrics* **76(2)**, 273–290.

[4] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19(6)**, 716–723. ISSN 0018-9286.

[5] Albert, R. (2007). Network Inference, Analysis, and Modeling in Systems Biology. *The Plant Cell Online* **19(11)**, 3327–3338. ISSN 1040-4651.

[6] Axelrod, R.; Hamilton, W. D. (1981). The evolution of cooperation. *science* **211(4489)**, 1390–1396.

[7] Azoulay, P.; Fons-Rosen, C.; Zivin, J. S. G. (2014). Does Science Advance One Funeral at a Time?

[8] Bani-Ahmad, S.; Cakmak, A.; Al-Hamdani, A.; Ozsoyoglu, G. (2005). Evaluating Score and Publication Similarity Functions in Digital Libraries. In: E. A. Fox; E. J. Neuhold; P. Premsmit; V. Wuwongse (eds.), *Digital Libraries: Implementing Strategies and Sharing Experiences*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 483–485. ISBN 978-3-540-32291-7.

[9] Barabási, A.-L.; Albert, R. (1999). Emergence of Scaling in Random Networks. *Science* **286(5439)**, 509–512. ISSN 0036-8075.

[10] Barrat, A.; Cattuto, C.; Colizza, V.; Pinton, J.; Van den Broeck, W.; Vespignani, A. (2008). High Resolution Dynamical Mapping of Social Interactions With Active RFID. *ArXiv e-prints* .

[11] Bascompte, J.; Jordano, P.; Melián, C. J.; Olesen, J. M. (2003). The nested assembly of plant–animal mutualistic networks. *Proceedings of the National Academy of Sciences* **100(16)**, 9383–9387.

[12] Benbya, H.; McKelvey, B. (2006). Toward a complexity theory of information systems development. *Information Technology & People* **19(1)**, 12–34.

[13] Bhattacharya, I.; Getoor, L. (2007). Collective Entity Resolution in Relational Data. *ACM Trans. Knowl. Discov. Data* **1(1)**. ISSN 1556-4681.

[14] Bianconi, G.; Barabási, A.-L. (2001). Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)* **54(4)**, 436.

[15] Bishop, C. M. (2006). Linear models for classification. In: *Pattern Recognition and Machine Learning*. Springer, pp. 179–224.

[16] Bloor, D. (2004). *Sociology of Scientific Knowledge*, Dordrecht: Springer Netherlands. ISBN 978-1-4020-1986-9, pp. 919–962.

[17] Boccaletti, S.; Bianconi, G.; Criado, R.; del Genio, C.; Gómez-Gardeñes, J.; Romance, M.; Sendiña-Nadal, I.; Wang, Z.; Zanin, M. (2014). The structure and dynamics of multilayer networks. *Physics Reports* **544(1)**, 1 – 122. ISSN 0370-1573. The structure and dynamics of multilayer networks.

[18] Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics Reports* **424(4–5)**, 175 – 308. ISSN 0370-1573.

[19] Bollen, J.; Van de Sompel, H.; Hagberg, A.; Chute, R. (2009). A Principal Component Analysis of 39 Scientific Impact Measures. *PLOS ONE* **4(6)**, 1–11.

[20] Bollobás, B. (1998). *Modern Graph Theory*, New York, NY: Springer New York, chap. Random Graphs. ISBN 978-1-4612-0619-4, pp. 215–252.

[21] Bolton, G. E.; Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American economic review* **90(1)**, 166–193.

[22] Börner, K.; Dall'Asta, L.; Ke, W.; Vespignani, A. (2005). Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams. *Complexity* **10(4)**, 57–67. ISSN 1099-0526.

[23] Boyack, K. W.; Klavans, R. (2010). Co□citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology* **61(12)**, 2389–2404.

[24] Boyack, K. W.; Klavans, R.; Börner, K. (2005). Mapping the backbone of science. *Scientometrics* **64(3)**, 351–374. ISSN 1588-2861.

[25] Brockmann, D.; Hufnagel, L.; Geisel, T. (2006). The scaling laws of human travel. *Nature* **439**.

[26] Brown, C. D.; Davis, H. T. (2006). Receiver operating characteristics curves and related decision measures: A tutorial. *Chemometrics and Intelligent Laboratory Systems* **80(1)**, 24 – 38. ISSN 0169-7439.

[27] Burrell, Q. L. (2005). Are "sleeping beauties" to be expected? *Scientometrics* **65(3)**, 381–389.

[28] Butts, C. T. (2009). Revisiting the foundations of network analysis. *science* **325(5939)**, 414–416.

[29] Byrd, R. H.; Lu, P.; Nocedal, J.; Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing* **16(5)**, 1190–1208.

[30] Börner, K.; Maru, J. T.; Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences* **101(suppl 1)**, 5266–5273.

[31] Börner, K.; Scharnhorst, A. (2009). Visual conceptualizations and models of science. *Journal of Informetrics* **3(3)**, 161 – 172. ISSN 1751-1577. Science of Science: Conceptualizations and Models of Science.

[32] Cabanac, G. (2011). Accuracy of inter-researcher similarity measures based on topical and social clues. *Scientometrics* **87(3)**, 597–620. ISSN 1588-2861.

[33] Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and program planning* **2(1)**, 67–90.

[34] Cartwright, D.; Harary, F. (1956). Structural balance: a generalization of Heider's theory. *Psychological review* **63(5)**, 277.

[35] Casiraghi, G. (2017). Multiplex Network Regression: How do relations drive interactions?

[36] Casiraghi, G.; Nanumyan, V.; Scholtes, I.; Schweitzer, F. (2016). Generalized Hypergeometric Ensembles: Statistical Hypothesis Testing in Complex Networks.

[37] Casiraghi, G.; Nanumyan, V.; Scholtes, I.; Schweitzer, F. (2017). From Relational Data to Graphs: Inferring Significant Links using Generalized Hypergeometric Ensembles. In: *International Conference on Social Informatics*. Springer, pp. 111–120.

[38] Cetina, K. (2009). *Epistemic Cultures: How the Sciences Make Knowledge*. Harvard University Press. ISBN 9780674039681.

[39] Cheng, E.; Grossman, J. W.; Lipman, M. J. (2003). Time-stamped graphs and their associated influence digraphs. *Discrete Applied Mathematics* **128(2)**, 317 – 335. ISSN 0166-218X.

[40] Chesson, J. (1976). A non-central multivariate hypergeometric distribution arising from biased sampling with application to selective predation. *Journal of Applied Probability* , 795–797.

[41] Cho, E.; Myers, S. A.; Leskovec, J. (2011). Friendship and Mobility: User Movement in Location-based Social Networks. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '11, New York, NY, USA: ACM, pp. 1082–1090. ISBN 978-1-4503-0813-7.

[42] Ciotti, V.; Bonaventura, M.; Nicosia, V.; Panzarasa, P.; Latora, V. (2016). Homophily and missing links in citation networks. *EPJ Data Science* **5(7)**.

[43] Cole, J. R.; Cole, S. (1972). The Ortega Hypothesis. *Science* **178(4059)**, 368–375. ISSN 0036-8075.

[44] Cole, S.; Simon, G. A.; *et al.* (1981). Chance and consensus in peer review. *Science* **214(4523)**, 881–886.

[45] Corcoran, C. D.; Senchaudhuri, P.; Mehta, C. R.; Patel, N. R. (2005). *Exact Inference for Categorical Data*, American Cancer Society. ISBN 9780470011812.

[46] Coscia, M.; Neffke, F. M. H. (2017). Network Backboning with Noisy Data. In: *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*. pp. 425–436.

[47] Crane, D. (1972). *Invisible Colleges; Diffusion of Knowledge in Scientific Communities*. University of Chicago Press. ISBN 9780226118574.

[48] Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology* **52(7)**, 558–569. ISSN 1532-2890.

[49] Day, T. E. (2015). The big consequences of small biases: A simulation of peer review. *Research Policy* **44(6)**, 1266 – 1270. ISSN 0048-7333.

[50] Doreian, P. (2004). Evolution of human signed networks. *Metodoloski zvezki* **1(2)**, 277.

[51] Doreian, P.; Mrvar, A. (2009). Partitioning signed social networks. *Social Networks* **31(1)**, 1 – 11. ISSN 0378-8733.

[52] Drenth, J. (1998). Multiple authorship: The contribution of senior authors. *JAMA* **280(3)**, 219–221.

[53] Duguid, P. (2005). "The Art of Knowing": Social and Tacit Dimensions of Knowledge and the Limits of the Community of Practice. *The Information Society* **21(2)**, 109–118.

[54] Dunham, Y.; Degner, J. (2010). Origins of intergroup bias: Developmental and social cognitive research on intergroup attitudes. *European Journal of Social Psychology* **40(4)**, 563–568. ISSN 1099-0992.

[55] Eagle, N.; Pentland, A. S.; Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* **106(36)**, 15274–15278.

[56] Eagle, N.; (Sandy) Pentland, A. (2006). Reality mining: sensing complex social systems. *Personal Ubiquitous Comput.* **10(4)**, 255–268. ISSN 1617-4909.

[57] van Eck, N. J.; Waltman, L. (2014). *Visualizing Bibliometric Networks*, Cham: Springer International Publishing. ISBN 978-3-319-10377-8, pp. 285–320.

[58] van Eck, N. J.; Waltman, L.; Noyons, E. C. M.; Buter, R. K. (2010). Automatic term identification for bibliometric mapping. *Scientometrics* **82(3)**, 581–596. ISSN 1588-2861.

[59] Eck, N. J. v.; Waltman, L. (2009). How to normalize cooccurrence data? An analysis of some well-known similarity measures. *Journal of the American Society for Information Science and Technology* **60(8)**, 1635–1651. ISSN 1532-2890.

[60] Egghe, L.; Rousseau, R. (2002). Co-citation, bibliographic coupling and a characterization of lattice citation networks. *Scientometrics* **55(3)**, 349–361.

[61] Ehrig, M.; Haase, P.; Hefke, M.; Stojanovic, N. (2005). Similarity for ontologies-a comprehensive framework. *ECIS 2005 Proceedings* , 127.

[62] Eppstein, D.; Galil, Z.; Italiano, G. F.; Nissenzweig, A. (1997). Sparsification—a Technique for Speeding Up Dynamic Graph Algorithms. *J. ACM* **44(5)**, 669–696. ISSN 0004-5411.

[63] Erdös, P.; Rényi, A. (1959). On random graphs I. *Publ. Math. Debrecen* **6**, 290–297.

[64] Expert, P.; Evans, T. S.; Blondel, V. D.; Lambiotte, R. (2011). Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences* **108(19)**, 7663–7668.

[65] Fang, F. C.; Casadevall, A. (2011). Retracted Science and the Retraction Index. *Infection and Immunity* **79(10)**, 3855–3859.

[66] Fiske, A. P. (1992). The four elementary forms of sociality: framework for a unified theory of social relations. *Psychological review* **99(4)**, 689.

[67] Fister Jr, I.; Fister, I.; Perc, M. (2016). Toward the discovery of citation cartels in citation networks. *Frontiers in Physics* **4**, 49.

[68] Fog, A. (2008). Calculation Methods for Wallenius' Noncentral Hypergeometric Distribution. *Communications in Statistics - Simulation and Computation* **37(2)**, 258–273. ISSN 0361-0918.

[69] Fog, A. (2008). Sampling Methods for Wallenius' and Fisher's Noncentral Hypergeometric Distributions. *Communications in Statistics - Simulation and Computation* **37(2)**, 241–257.

[70] Fortunato, S.; Bergstrom, C. T.; Börner, K.; Evans, J. A.; Helbing, D.; Milojević, S.; Petersen, A. M.; Radicchi, F.; Sinatra, R.; Uzzi, B.; Vespignani, A.; Waltman, L.; Wang, D.; Barabási, A.-L. (2018). Science of science. *Science* **359(6379)**. ISSN 0036-8075.

[71] Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social networks* **1(3)**, 215–239.

[72] Friedkin, N. E. (1991). Theoretical foundations for centrality measures. *American journal of Sociology* , 1478–1504.

[73] Frost, J. (2014). How Important Are Normal Residuals in Regression Analysis? `http://web.archive.org/web/20170909102236/http://blog.minitab.com:80/blog/adventures-in-statistics-2/how-important-are-normal-residuals-in-regression-analysis`. Accessed: 2018-03-10.

[74] Garas, A.; Tomasello, M. V.; Schweitzer, F. (2014). Newcomers vs. incumbents: How firms select their partners for R&D collaborations.

[75] Garcia, D.; Abisheva, A.; Schweighofer, S.; Serdült, U.; Schweitzer, F. (2015). Ideological and temporal components of network polarization in online political participatory media. *Policy & Internet* **7(1)**, 46–79.

[76] García, D.; Tanase, D. (2013). Measuring cultural dynamics through the eurovision song contest. *Advances in Complex Systems* **16(08)**, 1350037.

[77] Garlaschelli, D.; Loffredo, M. I. (2004). Patterns of Link Reciprocity in Directed Networks. *Phys. Rev. Lett.* **93**, 268701.

[78] Glattfelder, J. B.; Battiston, S. (2009). Backbone of complex networks of corporations: The flow of control. *Phys. Rev. E* **80**, 036104.

[79] Golosovsky, M. (2018). Preferential attachment mechanism of complex network growth: "rich-gets-richer" or "fit-gets-richer"? .

[80] Golosovsky, M.; Solomon, S. (2012). Stochastic Dynamical Model of a Growing Citation Network Based on a Self-Exciting Point Process. *Phys. Rev. Lett.* **109**, 098701.

[81] Golosovsky, M.; Solomon, S. (2017). Growing complex network of citations of scientific papers: Modeling and measurements. *Phys. Rev. E* **95**, 012324.

[82] Goodhart, C. A. (1984). Problems of monetary management: the UK experience. In: *Monetary Theory and Practice*, Springer. pp. 91–121.

[83] Goodreau, S. M.; Kitts, J. A.; Morris, M. (2009). Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography* **46(1)**, 103–125.

[84] Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American sociological review* , 161–178.

[85] Grady, D.; Thiemann, C.; Brockmann, D. (2012). Robust classification of salient links in complex networks. *Nature Communications* **3**, 864 EP –. Article.

[86] Guimerà, R.; Uzzi, B.; Spiro, J.; Amaral, L. A. N. (2005). Team Assembly Mechanisms Determine Collaboration Network Structure and Team Performance. *Science* **308(5722)**, 697–702. ISSN 0036-8075.

[87] Hagerstrand, T. (1966). Aspects of the spatial structure of social communication and the diffusion of information. *Papers in Regional Science* **16(1)**, 27–42.

[88] Harary, F.; *et al.* (1953). On the notion of balance of a signed graph. *The Michigan Mathematical Journal* **2(2)**, 143–146.

[89] Heider, F. (2013). *The psychology of interpersonal relations*. Psychology Press.

[90] Herrera, M.; Roberts, D. C.; Gulbahce, N. (2010). Mapping the Evolution of Scientific Fields. *PLoS ONE* **5(5)**, 1–6.

[91] Hofstede, G. (1984). *Culture's consequences: International differences in work-related values*, vol. 5. sage.

[92] Holland, P. W.; Laskey, K. B.; Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks* **5(2)**, 109 – 137. ISSN 0378-8733.

[93] Hosmer Jr, D. W.; Lemeshow, S.; Sturdivant, R. X. (2013). *Applied logistic regression*, vol. 398. John Wiley & Sons.

[94] Ingham, A. G.; Levinger, G.; Graves, J.; Peckham, V. (1974). The Ringelmann effect: Studies of group size and group performance. *Journal of Experimental Social Psychology* **10(4)**, 371 – 384. ISSN 0022-1031.

[95] Jaccard, P. (1912). The distribution of the flora in the Alpine zone. *New Phytologist* **11(2)**, 37–50. ISSN 1469-8137.

[96] Jacod, J.; Protter, P. (2004). *Probability Essentials*. Springer, Berlin, Heidelberg. ISBN 978-3-642-55682-1.

[97] Jain, A. K.; Dubes, R. C. (1988). Algorithms for clustering data .

[98] Jeong, H.; Néda, Z.; Barabási, A. L. (2003). Measuring preferential attachment in evolving networks. *EPL (Europhysics Letters)* **61(4)**, 567.

[99] Jones, E.; Oliphant, T.; Peterson, P.; *et al.* (2001–). SciPy: Open source scientific tools for Python. [Online; accessed <today>].

[100] Karrer, B.; Newman, M. E. J. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**, 016107.

[101] Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation* **14(1)**, 10–25.

[102] King, D. A. (2004). The scientific impact of nations. *Nature* **430**, 311 EP –.

[103] Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago: University of Chicago Press, xii, 210 pp.

[104] Lambiotte, R.; Panzarasa, P. (2009). Communities, knowledge creation, and information diffusion. *Journal of Informetrics* **3(3)**, 180–190.

[105] Larivière, V.; Archambault, È.; Gingras, Y. (2008). Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). *Journal of the American Society for Information Science and Technology* **59(2)**, 288–296. ISSN 1532-2890.

[106] Larivière, V.; Gingras, Y.; Archambault, É. (2006). Canadian collaboration networks: A comparative analysis of the natural sciences, social sciences and the humanities. *Scientometrics* **68(3)**, 519–533. ISSN 1588-2861.

[107] Lasserre, J.; Bishop, C.; Bernardo, J.; Bayarri, M.; Berger, J.; Dawid, A.; Heckerman, D.; Smith, A.; West, M. (2007). *Generative or Discriminative? Getting the Best of Both Worlds*, Oxford University Press, vol. 8. ISBN 9780199214655, pp. 3–24.

[108] Leonard, D.; Sensiper, S. (1998). The Role of Tacit Knowledge in Group Innovation. *California Management Review* **40(3)**, 112–132. ISSN 0008-1256.

[109] Leskovec, J.; Backstrom, L.; Kumar, R.; Tomkins, A. (2008). Microscopic Evolution of Social Networks. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '08, New York, NY, USA: ACM, pp. 462–470. ISBN 978-1-60558-193-4.

[110] Leskovec, J.; Huttenlocher, D.; Kleinberg, J. (2010). 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: ACM, pp. 641–650. ISBN 978-1-60558-799-8.

[111] Leskovec, J.; Huttenlocher, D.; Kleinberg, J. (2010). Signed Networks in Social Media. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '10, New York, NY, USA: ACM, pp. 1361–1370. ISBN 978-1-60558-929-9.

[112] Leydesdorff, L.; Carley, S.; Rafols, I. (2013). Global maps of science based on the new Web-of-Science categories. *Scientometrics* **94(2)**, 589–593. ISSN 1588-2861.

[113] Leydesdorff, L.; Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology* **60(2)**, 348–362. ISSN 1532-2890.

[114] Li, G.-C.; Lai, R.; D'Amour, A.; Doolin, D. M.; Sun, Y.; Torvik, V. I.; Yu, A. Z.; Fleming, L. (2014). Disambiguation and co-authorship networks of the U.S. patent inventor database (1975–2010). *Research Policy* **43(6)**, 941 – 955. ISSN 0048-7333.

[115] Lü, L.; Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications* **390(6)**, 1150 – 1170. ISSN 0378-4371.

[116] Ma, A.; Mondragón, R. J.; Latora, V. (2015). Anatomy of funded research in science. *Proceedings of the National Academy of Sciences* **112(48)**, 14760–14765.

[117] Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India 2* **1**, 49–55.

[118] Maniu, S.; Cautis, B.; Abdessalem, T. (2011). Building a Signed Network from Interactions in Wikipedia. In: *Databases and Social Networks*. DBSocial '11, New York, NY, USA: ACM, pp. 19–24. ISBN 978-1-4503-0650-8.

[119] Marshakova, I. V. (1973). System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2-Informatsionnye Protsessy I Sistemy* **(6)**, 3–8.

[120] Martin; Delvenne, J.-C.; Schaub, M. T.; Lambiotte, R. (2017). Different approaches to community detection.

[121] McFadden, D.; *et al.* (1973). Conditional logit analysis of qualitative choice behavior .

[122] McPherson, M.; Smith-Lovin, L.; Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology* **27(1)**, 415–444.

[123] McPherson, M.; Smith-Lovin, L.; Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* **27(1)**, 415–444.

[124] Medo, M. c. v. (2014). Statistical validation of high-dimensional models of growing networks. *Phys. Rev. E* **89**, 032801.

[125] Medo, M. c. v.; Cimini, G.; Gualdi, S. (2011). Temporal Effects in the Growth of Networks. *Phys. Rev. Lett.* **107**, 238701.

[126] Merton, R. K. (1968). The Matthew Effect in Science. *Science* **159(3810)**, 56–63. ISSN 0036-8075.

[127] Middendorf, M.; Ziv, E.; Wiggins, C. H. (2005). Inferring network mechanisms: The Drosophila melanogaster protein interaction network. *Proceedings of the National Academy of Sciences of the United States of America* **102(9)**, 3192–3197.

[128] Milojević, S. (2014). Principles of scientific research team formation and evolution. *Proceedings of the National Academy of Sciences* **111(11)**, 3984–3989. ISSN 0027-8424.

[129] Molloy, M.; Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms* **6(2-3)**, 161–180. ISSN 1098-2418.

[130] Moreno, J.; Jennings, H. (1934). *Who shall survive?: A new approach to the problem of human interrelations*. Nervous and mental disease monograph series, Nervous and mental disease publishing co.

[131] Mucha, P. J.; Richardson, T.; Macon, K.; Porter, M. A.; Onnela, J.-P. (2010). Community Structure in Time-Dependent, Multiscale, and Multiplex Networks. *Science* **328(5980)**, 876–878. ISSN 0036-8075.

[132] Namata, G. M.; London, B.; Getoor, L. (2015). Collective Graph Identification. *ACM Transactions on Knowledge Discovery from Data* **10(3)**, 25:1–25:36.

[133] Nanumyan, V. (2014). *Master Equations for Heterogeneous Evolution of Interdependent Networks*. Master's thesis, ETH Zürich.

[134] Nanumyan, V.; Garas, A.; Schweitzer, F. (2015). The Network of Counterparty Risk: Analysing Correlations in OTC Derivatives. *PLOS ONE* **10(9)**, 1–23.

[135] Newman, M. (2009). The first-mover advantage in scientific publication. *EPL (Europhysics Letters)* **86(6)**, 68001.

[136] Newman, M. (2010). *Networks: an introduction*. Oxford University Press, Oxford.

[137] Newman, M. E.; Clauset, A. (2016). Structure and inference in annotated networks. *Nature communications* **7**, 11863.

[138] Newman, M. E. J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences* **98(2)**, 404–409.

[139] Newman, M. E. J. (2003). The Structure and Function of Complex Networks. *SIAM Review* **45(2)**, 167–256.

[140] Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences* **103(23)**, 8577–8582. ISSN 0027-8424.

[141] Newman, M. E. J. (2018). Network reconstruction and error estimation with noisy network data.

[142] Newman, M. E. J.; Peixoto, T. P. (2015). Generalized Communities in Networks. *Phys. Rev. Lett.* **115**, 088701.

[143] Nicosia, V.; Bianconi, G.; Latora, V.; Barthelemy, M. (2013). Growing Multiplex Networks. *Phys. Rev. Lett.* **111**, 058701.

[144] Noulas, A.; Scellato, S.; Lambiotte, R.; Pontil, M.; Mascolo, C. (2012). A tale of many cities: universal patterns in human urban mobility. *PloS one* **7(5)**, e37027.

[145] Otte, E.; Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of information Science* **28(6)**, 441–453.

[146] Page, L.; Brin, S.; Motwani, R.; Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66*, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

[147] Pahl, R. E. (1988). Some remarks on informal work, social polarization and the social structure. *International Journal of Urban and Regional Research* **12(2)**, 247–267.

[148] Palchykov, V.; Gemmetto, V.; Boyarsky, A.; Garlaschelli, D. (2016). Ground truth? Concept-based communities versus the external classification of physics manuscripts. *EPJ Data Science* **5(1)**, 28. ISSN 2193-1127.

[149] Parolo, P. D. B.; Pan, R. K.; Ghosh, R.; Huberman, B. A.; Kaski, K.; Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics* **9(4)**, 734 – 745. ISSN 1751-1577.

[150] Peel, L.; Larremore, D. B.; Clauset, A. (2017). The ground truth about metadata and community detection in networks. *Science Advances* **3(5)**.

[151] Peixoto, T. P. (2014). Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Phys. Rev. E* **89**, 012804.

[152] Peters, D. P.; Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behavioral and Brain Sciences* **5(02)**, 187–195.

[153] Pfitzner, R.; Scholtes, I.; Garas, A.; Tessone, C. J.; Schweitzer, F. (2013). Betweenness Preference: Quantifying Correlations in the Topological Dynamics of Temporal Networks. *Phys. Rev. Lett.* **110**, 198701.

[154] Pigliucci, M. (2009). The end of theory in science? *EMBO reports* **10(6)**, 534–534. ISSN 1469-221X.

[155] Polanyi, M. (1966). *The tacit dimension*. University of Chicago Press.

[156] Pravdić, N.; Oluić-Vuković, V. (1986). Dual approach to multiple authorship in the study of collaboration/scientific output relationship. *Scientometrics* **10(5)**, 259–280. ISSN 1588-2861.

[157] Price, D. D. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* **27(5)**, 292–306. ISSN 1097-4571.

[158] Price, D. J. d. S. (1963). *Little science, big science*. New York: Columbia Univ. Press. ISBN 0231085621 9780231085625.

[159] Price, D. J. D. S. (1965). Networks of Scientific Papers. *Science* **149(3683)**, 510–515. ISSN 00368075, 10959203.

[160] Pyka, A.; Scharnhorst, A. (2009). *Introduction: Network Perspectives on Innovations: Innovative Networks – Network Innovation*, Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-92267-4, pp. 1–16.

[161] Radicchi, F.; Fortunato, S.; Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences* **105(45)**, 17268–17272. ISSN 0027-8424.

[162] Ren, Z.-M.; Mariani, M. S.; Zhang, Y.-C.; Medo, M. c. v. (2018). Randomizing growing networks with a time-respecting null model. *Phys. Rev. E* **97**, 052311.

[163] Robins, G.; Pattison, P.; Kalish, Y.; Lusher, D. (2007). An introduction to exponential random graph (p*) models for social networks. *Social networks* **29(2)**, 173–191.

[164] Samarin, M. (2016). *Modelling co-locations in human mobility*. Master's thesis, ETH Zürich.

[165] Sarigol, E.; Pfitzner, R.; Scholtes, I.; Garas, A.; Schweitzer, F. (2014). Predicting Scientific Success Based on Coauthorship Networks. *EPJ Data Science* **3**, 9.

[166] Sawyer, K. (1984). Multiple hypothesis testing. *Journal of the Royal Statistical Society. Series B (Methodological)* , 419–424.

[167] Scharnhorst, A. (2001). *Constructing Knowledge Landscapes Within the Framework of Geometrically Oriented Evolutionary Theories*, Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-56585-4, pp. 505–515.

[168] Schneider, C. M.; Belik, V.; Couronné, T.; Smoreda, Z.; González, M. C. (2013). Unravelling daily human mobility motifs. *Journal of The Royal Society Interface* **10(84)**. ISSN 1742-5689.

[169] Scholtes, I. (2017). When is a Network a Network? Multi-Order Graphical Model Selection in Pathways and Temporal Networks. *ArXiv e-prints* .

[170] Scholtes, I.; Mavrodiev, P.; Schweitzer, F. (2016). From Aristotle to Ringelmann: a large-scale analysis of team productivity and coordination in Open Source Software projects. *Empirical Software Engineering* **21(2)**, 642–683. ISSN 1573-7616.

[171] Scholtes, I.; Pfitzner, R.; Schweitzer, F. (2014). The Social Dimension of Information Ranking: A Discussion of Research Challenges and Approaches. In: *Socioinformatics - The Social Impact of Interactions between Humans and IT*. Springer Proceedings in Complexity, Springer International Publishing, pp. 45–61. ISBN 978-3-319-09377-2.

[172] Scholtes, I.; Wider, N.; Pfitzner, R.; Garas, A.; Tessone, C. J.; Schweitzer, F. (2014). Causality-driven slow-down and speed-up of diffusion in non-Markovian temporal networks. *Nature Communications* **5**, 5024 EP –. Article.

[173] Schwarz, G. (1978). Estimating the Dimension of a Model. *Ann. Statist.* **6(2)**, 461–464.

[174] Schweighofer, S. (2018). *Affective, Cognitive, and Social Identity Related Factors of Political Polarization*. Doctoral dissertation, ETH Zurich. Diss. ETH No. 24797.

[175] Schweitzer, F. (2003). *Brownian Agents and Active Particles: Collective Dynamics in the Natural and Social Sciences*. Physics and astronomy online library, Springer Berlin Heidelberg. ISBN 9783540439387.

[176] Schweitzer, F.; Nanumyan, V.; Tessone, C. J.; Xia, X. (2014). How do OSS projects change in number and size? A large-scale analysis to test a model of project growth. *Advances in Complex Systems* **17(07n08)**, 1550008.

[177] Serrano, M. Á.; Boguñá, M.; Vespignani, A. (2009). Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences* **106(16)**, 6483–6488. ISSN 0027-8424.

[178] Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual review of psychology* **46(1)**, 561–584.

[179] Simmel, G. (1996). Der Streit. In: *Konflikttheorien*, Springer. pp. 240–262.

[180] Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika* **42(3-4)**, 425–440.

[181] Sinatra, R.; Wang, D.; Deville, P.; Song, C.; Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science* **354(6312)**. ISSN 0036-8075.

[182] Singer, P.; Helic, D.; Hotho, A.; Strohmaier, M. (2015). HypTrails: A Bayesian Approach for Comparing Hypotheses About Human Trails on the Web. In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15, Republic and Canton of Geneva, Switzerland: International World Wide Web Conferences Steering Committee, pp. 1003–1013. ISBN 978-1-4503-3469-3.

[183] Skyrms, B.; Pemantle, R. (2000). A dynamic model of social network formation. *Proceedings of the National Academy of Sciences* **97(16)**, 9340–9346.

[184] Slater, P. B. (2009). A two-stage algorithm for extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences* **106(26)**, E66–E66.

[185] Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the Association for Information Science and Technology* **24(4)**, 265–269.

[186] Snijders, T. A.; Pattison, P. E.; Robins, G. L.; Handcock, M. S. (2006). New specifications for exponential random graph models. *Sociological methodology* **36(1)**, 99–153.

[187] Song, C.; Koren, T.; Wang, P.; Barabási, A.-L. (2010). Modelling the scaling properties of human mobility. *Nature Physics* **6**, 818 EP –. Article.

[188] Sprinthall, R. (2012). *Basic Statistical Analysis*. Pearson Allyn & Bacon. ISBN 9780205052172.

[189] Sprinzak, E.; Sattath, S.; Margalit, H. (2003). How reliable are experimental protein–protein interaction data? *Journal of molecular biology* **327(5)**, 919–923.

[190] Squartini, T.; Garlaschelli, D. (2017). *Maximum-Entropy Networks: Pattern Detection, Network Reconstruction and Graph Combinatorics*. SpringerBriefs in Complexity, Springer International Publishing. ISBN 9783319694382.

[191] Su, H.-N.; Lee, P.-C. (2010). Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in Technology Foresight. *Scientometrics* **85(1)**, 65–79. ISSN 1588-2861.

[192] Sutcliffe, A.; Dunbar, R.; Binder, J.; Arrow, H. (2012). Relationships and the social brain: integrating psychological and evolutionary perspectives. *British journal of psychology* **103(2)**, 149–168.

[193] Sutton, J. (1997). Gibrat's Legacy. *Journal of Economic Literature* **35(1)**, 40–59.

[194] Tomasello, M. V.; Tessone, C. J.; Schweitzer, F. (2016). A model of dynamic rewiring and knowledge exchange in R&D networks. *Advances in Complex Systems* **19(1-2)**.

[195] Tomasello, M. V.; Vaccario, G.; Schweitzer, F. (2017). Data-driven modeling of collaboration networks: a cross-domain analysis.

[196] Toole, J. L.; Herrera-Yaqüe, C.; Schneider, C. M.; González, M. C. (2015). Coupling human mobility and social ties. *Journal of The Royal Society Interface* **12(105)**. ISSN 1742-5689.

[197] Torvik, V. I.; Weeber, M.; Swanson, D. R.; Smalheiser, N. R. (2005). A probabilistic similarity metric for Medline records: A model for author name disambiguation. *Journal of the American Society for Information Science and Technology* **56(2)**, 140–158. ISSN 1532-2890.

[198] Tuomi, I. (1999). Data Is More than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory. *Journal of Management Information Systems* **16(3)**, 103–117.

[199] Uzzi, B.; Mukherjee, S.; Stringer, M.; Jones, B. (2013). Atypical Combinations and Scientific Impact. *Science* **342(6157)**, 468–472. ISSN 0036-8075.

[200] Vaccario, G.; Medo, M.; Wider, N.; Mariani, M. S. (2017). Quantifying and suppressing ranking bias in a large citation network. *Journal of Informetrics* **11(3)**, 766–782.

[201] Van Raan, A. F. (2004). Sleeping beauties in science. *Scientometrics* **59(3)**, 467–472.

[202] Wagner, C. S.; Leydesdorff, L. (2005). Network structure, self-organization, and the growth of international collaboration in science. *Research Policy* **34(10)**, 1608 – 1618. ISSN 0048-7333.

[203] Wallenius, K. T. (1963). *Biased Sampling: the Noncentral Hypergeometric Probability Distribution*. Ph.d. thesis, Stanford University.

[204] Wang, D.; Song, C.; Barabási, A.-L. (2013). Quantifying Long-Term Scientific Impact. *Science* **342(6154)**, 127–132. ISSN 0036-8075.

[205] Wang, M.; Yu, G.; Yu, D. (2008). Measuring the preferential attachment mechanism in citation networks. *Physica A: Statistical Mechanics and its Applications* **387(18)**, 4692–4698.

[206] Wang, M.; Yu, G.; Yu, D. (2009). Effect of the age of papers on the preferential attachment in citation networks. *Physica A: Statistical Mechanics and its Applications* **388(19)**, 4273–4276.

[207] White, H. D.; Griffith, B. C. (1981). Author cocitation: A literature measure of intellectual structure. *Journal of the Association for Information Science and Technology* **32(3)**, 163–171.

[208] Wider, N.; Garas, A.; Scholtes, I.; Schweitzer, F. (2016). *Interconnected Networks*, Cham: Springer International Publishing, chap. An Ensemble Perspective on Multi-layer Networks. ISBN 978-3-319-23947-7, pp. 37–59.

[209] Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Ann. Math. Statist.* **9(1)**, 60–62.

[210] Winkler, W. E. (1999). *The State of Record Linkage and Current Research Problems*. *Tech. rep.*, Statistical Research Division, U.S. Census Bureau.

[211] Wu, F.; Huberman, B. A. (2007). Novelty and collective attention. *Proceedings of the National Academy of Sciences* **104(45)**, 17599–17601. ISSN 0027-8424.

[212] Yan, E.; Ding, Y. (2012). Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other. *Journal of the American Society for Information Science and Technology* **63(7)**, 1313–1326. ISSN 1532-2890.

[213] Yang, S.-H.; Smola, A. J.; Long, B.; Zha, H.; Chang, Y. (2012). Friend or Frenemy?: Predicting Signed Ties in Social Networks. In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '12, New York, NY, USA: ACM, pp. 555–564. ISBN 978-1-4503-1472-5.

[214] Yitzhaki, M. (1994). Relation of title length of journal articles to number of authors. *Scientometrics* **30(1)**, 321–332.

[215] Zachary, W. W. (1977). An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research* **Vol. 33(No. 4)**, 452–473.

[216] Zhao, D.; Strotmann, A. (2008). Evolution of research activities and intellectual influences in information science 1996–2005: Introducing author bibliographic-coupling analysis. *Journal of the Association for Information Science and Technology* **59(13)**, 2070–2086.

[217] Zhou, T.; Ren, J.; Medo, M.; Zhang, Y.-C. (2007). Bipartite network projection and personal recommendation. *Physical Review E* **76(4)**, 046115.

[218] Zingg, C. (2017). *Authors' influence on citation rate dynamics*. Master's thesis, ETH Zürich.

[219] Zins, C. (2007). Conceptual approaches for defining data, information, and knowledge. *Journal of the American Society for Information Science and Technology* **58(4)**, 479–493. ISSN 1532-2890.

[220] Zins, C.; Santos, P. L. (2011). Mapping the knowledge covered by library classification systems. *Journal of the American Society for Information Science and Technology* **62(5)**, 877–901. ISSN 1532-2890.