Department of Computer Science

Analysis of Cumulative and Temporal Patterns in Science

Pietro della Briotta Parolo

Aalto University

DOCTORAL
DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 226/2017

Analysis of Cumulative and Temporal Patterns in Science

Pietro della Briotta Parolo

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall TU1 of the school on 4 December 2017 at 13.

Aalto University
School of Science
Department of Computer Science

Supervising professor
Professor Kimmo Kaski, Aalto University, Finland

Thesis advisors
Assistant Professor Mikko Kivelä, Aalto University, Finland
Professor Santo Fortunato, Indiana University, USA

Preliminary examiners
Assistant Professor Alexander M., University of California Merced, USA

Professor Anxo Sánchez, Universidad Carlos III de Madrid, Spain

Opponents
Assistant Professor Márton Karsai, École Normale Supérieure de Lyon, France

Aalto University publication series
DOCTORAL DISSERTATIONS 226/2017

ISBN 978-952-60-7725-3 (printed)
ISBN 978-952-60-7726-0 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-7726-0

Unigrafia Oy
Helsinki 2017

Finland

NORDIC ECOLABEL
441 697
Printed matter

A? Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Pietro della Briotta Parolo

Name of the doctoral dissertation
Analysis of Cumulative and Temporal Patterns in Science

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 226/2017

Field of research Science of Science

Language English

□ **Monograph** ☒ **Article dissertation** □ **Essay dissertation**

Abstract

The goal of science has always been to investigate the world and its phenomena, by collecting data from all possible events that take place around us, breaking them down into their most simple elements and trying to come up with models able to explain and predict the outcome of these events. For centuries, the primary focus of science was mainly on natural events, but as the new technologies allowed to gather data from human interactions, it was natural for scientists to use this new information in order to apply the same logic to social systems, including science itself.

Since the late 19th century, when the first modern scientific journals were published, science has seen a constant rise in both its size and productivity, thanks to the standardization of research practices and the building of an international community that actively helps to push forward the limits of human knowledge. As science itself went from being a purely intellectual endeavor to a complex social, economical and political system, it is no surprise that a lot of attention has been dedicate in recent years to the study of the underlying mechanisms of science, aided by the explosion of means of communication that allow collaborations and exchange of information at instant speed across the globe, leaving behind digital traces that provide valuable data to study. The continuous exponential growth of science however, causes also difficulties in analyzing objectively the patterns and statistics that scientific data can reveal: for example a paper from the early 20th century would rarely get more than 100 citations, while now it is not uncommon for publications to pass the 10 thousand citation mark.

This thesis follows these attempts in trying to grasp how science works, by investigating the connections, i.e. citations, that exists between scientific publications and how these connections create structures and patterns. It shows that typical patterns in citation count and diffusion of information between fields is heavily influenced by the rate of growth of science, thus suggesting to use the number of publications as a better measure of time. It shows that there is a lag between breakthrough discoveries and the time when they are recognized, thus suggesting that we might be either running out of discoveries or rather having too much of them, in either case an extreme phenomenon. It shows that the community of publications which builds around an original successful paper has a typical life cycle, with an initial clustering, followed by an inevitable breaking down. Finally, it offers a new way of quantifying the impact of publications across time based on their cumulative impact on the overall corpus of scientific material.

Keywords science of science, scientometrics

ISBN (printed) 978-952-60-7725-3 **ISBN (pdf)** 978-952-60-7726-0

ISSN-L 1799-4934 **ISSN (printed)** 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2017

Pages 100 **urn** http://urn.fi/URN:ISBN:978-952-60-7726-0

Preface

This thesis is the product of the work started at BECS in winter 2013 and eventually completed at the CS department in the summer of 2017.

Looking back at these 4 years it's an interesting and exciting dive into a lot of moments that have influenced heavily my life both privately as well as professionally. Moving to Finland to pursue my academic passion was not just a big step forward in my career, but also a big leap into a new world of which I admit I knew very little about beforehand. The first days were therefore often as exciting as well as confusing and I need to thank the whole staff working at BECS at the time for keeping an eye on me. In particular I'd like to mention Lauri Kovanen, who showed me the ways of the Python, Arnab Chatterjee and Marija Mitrović, with whom I often discussed about Science in general, Raj Kumar Pan and Richard Darst, who had the patience to help me with often small technical issues that I was not able to overcome on my own. Also, a huge thanks goes to my doctoral office mate, Darko Hric, with whom I shared all the ups and downs of this adventure.

Leaving BECS was a huge change from many aspects. Most of the senior staff in the meantime had decided to leave Aalto, ultimately leading to the departure also of my supervisor, Prof. Santo Fortunato, whom I thank for guiding me and making me understand the pros and cons of life in academia. This period was extremely hectic, especially also with the organization of the iccss conference, which unfortunately does not show up as part of this dissertation, but which is one of my proudest contribution to Science, given the amount of work required and the quality of the final product, for which Prof. Fortunato. This period culminated with yet another move to a new office within the CS department, which represented the final stage of my stay at Aalto, now with the supervision of Dr. Mikko Kivelä and Prof. Kimmo Kaski, to whom I owe a huge amount of gratitude for "adopting" me in the final stage of my PhD and allowing me to conclude a project that turned out to be both exciting and productive. Prof. Kaski was not only an excellent scientific guide, but also a key figure in terms of supporting me, both personally and professionally, until the very end of my experience at Aalto.

Finally, a big thanks to all those who have had the unfortunate experience of being around me during my studies, but withouth whose support I wouldn’t have been able to get this far.

Espoo, November 7, 2017,

Pietro della Briotta Parolo

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Francesco Becattini, Arnab Chatterjee, Santo Fortunato, Marija Mitrović, Raj Kumar Pan, Pietro Della Briotta Parolo. The Nobel Prize delay . *Physics Today*, DOI:10.1063/PT.5.2012, May 2014.

II Pietro Della Briotta Parolo, Raj Kumar Pan, Rumi Ghosh Bernardo A. Huberman, Kimmo Kaski, Santo Fortunato. Attention Decay in Sciene. *Journal of Informetrics*, Volume 9, Issue 4, Pages 734–745, October 2015.

III Pietro Della Briotta Parolo, Santo Fortunato. Uncovering the Dynamics of Ego Networks of Scientific Gems. *preprint*, submitted to peer review, January 2017.

IV Pietro Della Briotta Parolo, Mikko Kivelä, Kimmo Kaski. On the Shoulders of Giants: tracking the cumulative knowledge spreading in citation networks. *preprint*, submitted to peer review, June 2017.

Author's Contribution

Publication I: "The Nobel Prize delay "

The author contributed to the collection of the data.

Publication II: "Attention Decay in Sciene"

The author carried out most of the analysis. Primary writer of the article.

Publication III: "Uncovering the Dynamics of Ego Networks of Scientific Gems"

The author implemented the analysis. Major role in writing the article.

Publication IV: "On the Shoulders of Giants: tracking the cumulative knowledge spreading in citation networks"

The author implemented the analysis. Major role in writing the article.

1. Introduction

1.1 Science of Science

The underlying driving force of science has always been to start from empirical evidence in order to gain information about the structure of the phenomena taking place around us. With such a pursuit in mind it was just a matter of time until science would start investigating *itself*. The moment came in the 60s, when the first bibliographic efforts required to improve the searchability of scientific material took place in the form of a search for proper indexing [1, 2] and therefore allowing, for the first time, to analyze the published material as its own data set. With only few previous works being carried out [3], the historical breakthrough in the field of science of science came with De Solla Price's work *Networks of Scientific Papers* [4]. De Solla's publication not only was one of the first to directly tackle the pattern of bibliographical references, but it also introduced key concepts for the development of the field, starting from the need to analyze it in its topological structure as a network. Figs.1.1 and 1.2 show the earliest attempts of representing citation data as a network, even though the theory behind network science was still in its earliest stages.

What is most striking however, is that already in its origins, the study of the scientific production has required an analysis of science *as a whole* and *in time*. These key features are intrinsic properties of the entire scientific production, since it is in the nature of science to build one's work on the top of previous ones, therefore adding a temporal dimension to its development, as new discoveries and breakthroughs appear and link themselves to older ones. Since that seminal paper, the whole world, as well as the scientific one, has seen an amazing rise in technological possibilities, which have affected heavily the opportunities for collaborations, allowing people, as well as ideas, to move freely across the globe.

These conditions, along with an improvement in the economies in the post War era, has allowed science to grow at an amazing rate [5]. The amount of information generated by science has been growing exponentially at a rate close to a 4% growth *each year* in the last decades as shown in Fig.1.3. Scientists

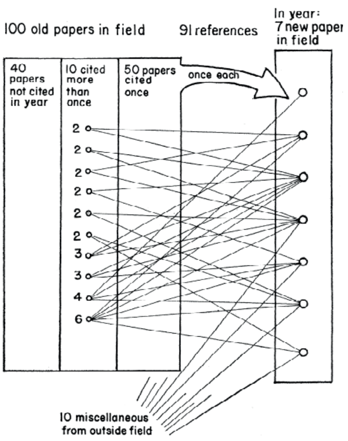

Figure 1.1. Representation of citations as a network structure. Figure adapted from [4] with permission of The American Association for the Advancement of Science.

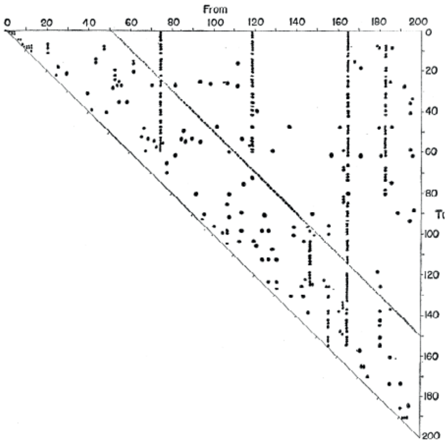

Figure 1.2. Representation of citations as an adjacency matrix. Figure adapted from [4] with permission of The American Association for the Advancement of Science.

are constantly dealing with the necessity to retrieve the latest results from their fields, which are also growing at a fast rate; in such framework the ability to focus on the most relevant works becomes a key aspect. However, need for constant update requires to shift one's attention towards more recent scientific results, gradually discarding older ones.

The same applies in the other direction, with scientists trying to get their latest publication known as much as possible, in order to gather *attention* on their latest results. Therefore, scientists are actors in a market where the ability of reaching popularity in terms of scientific productivity has become a dominating aspect, implying that scientists/groups/institutions are all competing for attention in a market where the allocation thereof is structurally limited by one's ability to store information regarding all scientific results published in the past.

1.2 Scope of the Thesis

This thesis focus mainly on this temporal and cumulative aspect, investigating the changes that science has undergone in time due to its constantly changing nature. Chapter 2 talks about the study of citation patterns, with their properties, biases and attempts at modeling them. Chapter 3 introduces the basic concepts of network theory and how these concepts have been used to analyze the social and collaborative structure of science. Chapter 4 talks about the efforts in trying to determine the quality of scientific publications by the development of *metrics*. Finally, Chapter 5 summarizes the content of Publications I-IV and

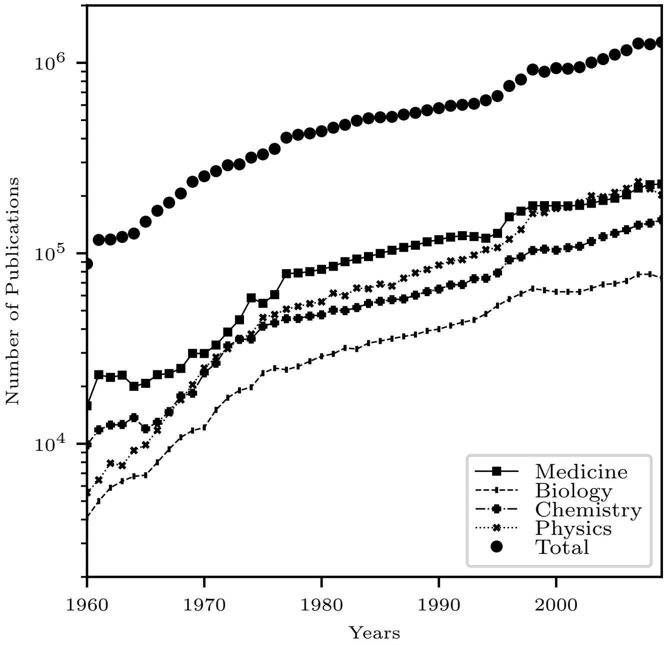

Figure 1.3. Growth of publication in science and for a selected number of fields based on our ISI dataset of over 50 million publications and 600 million citations. The rate of growth can be well approximated by an exponential curve. Figure adapted from Publication II.

discusses briefly how they contribute to the field of Science of Science.

2. Scientific Citations and Their Patterns

"*If I have seen further, it is by standing on the shoulders of giants*". This famous quote by Sir Isaac Newton summarizes perfectly the moral obligation of a scientist to acknowledge the contribution of previous works to their own. Newton was perfectly aware that his groundbreaking discoveries would have been impossible without the fundamental work done by previous scientists, from Aristotle to Galileo and Kepler, covering centuries, if not millennia of scientific and philosophical endeavours. While the recognition of the work done by predecessors at the times of Newton was done primarily by mentioning the names in the text or in private correspondence (as was the quote mentioned before) as a form of intellectual courtesy, in modern times it has taken the form in scientific journals of a moral obligation based on an agreed voluntary scheme and is considered as a fundamental part of good scientific practice, while for patents it even has a legal side, with previous patents being cited in order to be able to clarify how the new patent differs substantially from previously similar ones. Furthermore, due to the limited space available in a text, along with the gradual process that turns recent discoveries into common knowledge, the publications mentioned in the reference lists represent an extremely careful and precise process of selection of a very limited number of works among thousands, if not millions, of related works published in recent times.

As the results in aging literature are slowly assimilated as basic findings, scientists move on to newer results as the basis of their works, thus implicitly determining when a groundbreaking result becomes obsolete, as more impelling results require their attention. Just like Newton chose to acknowledge Galileo for a few selected results, but ignoring to do to the same with Pythagoras and his extensively used theorem, a recent paper in Quantum Physics will hardly mention any of the works of Einstein's Annus Mirabilis even though they are the very foundation on which its work is based on, since their results are now accepted as being universally known and do not need to be individually addressed anymore.

It is for these reasons that ever since the early times of scientometrics, a lot of attention has been given to the analysis of the individual performance of a single publication in terms of citations. A simple citation count is a superficial yet quantitative evaluation of the success of a paper and is deemed sufficient by some to be able to compare and rank publications as well as scientists. However, the aforementioned process of obsolesce in science adds a dimension which has been described as an *attention economy* [6] in which authors are aware that they have a limited amount of time to gather attention (i.e. citations) and therefore compete against each other in order to obtain the maximum attention available.

Such complex aspects that lead to the selection of the cited material has been the source of even more interest into the citation patterns as well as statistics of citation counts across disciplines, countries and through time. This chapter will go through the most relevant works that have investigated the citation patterns in science, looking at the basic properties in citation habits and with a summary of the most interesting attempts at modeling mathematically the citation patterns of scientists.

2.1 Citation distributions

One of the earliest questions that scientometrics tried to answer already with de Solla's seminal paper [4] has been: *What is the functional form of the distributions of citations?*. In particular, since the average value of citations gathered is bound to be structurally low as its value is linked to the finite number of references available, the interest was in the tail of the distributions, that is what are the citation values and patterns for the few exceptional publications capable to gather a number of citations that span over multiple orders of magnitude. De Solla claimed, based on his limited data, that the functional form was power law like, with the number of papers with c citations behaving like $N(c) \propto c^{-\alpha}$, with an estimate of $\alpha \in [2.5, 3.6]$.

For a long time, no one looked further into the claim with only Laherrère and Sornette in 1998 [7] suggesting a generic stretched exponential form for the citation distribution of *authors*. It was only in 1998 that S. Redner tackled the topic in a systematic way [8]. It is important to notice that such analysis was possible to be carried out mainly thanks to the availability of a properly catalogued data set of scientific publications. By using two large data sets (700 thousand papers obtained from the Institute for Scientific Information (ISI) and 24 thousand papers from Physical Review D) combining for more than 7 million citations, the author was able, for the first time, to carry out a thorough computational statistical analysis of citation distributions. The results offered an interesting and, to a certain extent, worrisome insight of the relative popularity of scientific publications: almost half of the papers failed to gather any citation at all from publication date to the time of the study, with 80% of the publications gathering 10 citations or less. Even though also de Solla noticed a huge amount of uncited papers, Redner was able to confirm the pattern also for a larger and more significant data set. The author concluded that a final evaluation of the functional form of the citation distribution cannot be thoroughly computed as the tail of the distribution has not reached its final state, as the highly cited papers are still gathering citations. He also pointed out how a few highly cited papers can affect the higher-order moments of the distributions, thus making the task even harder. However, Redner succeeded in gathering some indirect measurement through a Zipf plot [9], providing evidence of a power law behaviour with $\alpha \approx 3$, compatible with de Solla's findings. Furthermore, the author concluded with what can be considered the *cookbook* for future attempts at modeling the citation mechanism: a short memory (or Myopia) and the "rich get richer" kind of mechanism that was introduced by de Solla himself in 1976 [10]. The latter would become a massive topic starting from the following year, with Barabási's work on scaling in random networks [11] which managed to mathematically justify the power law distribution of citations.

Despite the case seeming to be settled, it was Redner himself in 2005 who challenged his own previous findings [12]. In his later work, the author looked deeper in the PR data set, this time expanded to over 300 thousand papers from July 1893 through June 2003, suggesting that a log-normal distribution better describes the data.

A somewhat conclusive result in the discussion of the form of citation distributions came in 2008 with the work of Radicchi et al. [13] who found strong evidence for a lognormal distribution for the citation distribution of scientific publications and furthermore managed to discover universal properties in the citation distribution across disciplines as different fields have. In their paper, the authors show how the citation distributions across fields, despite being apparently extremely different quantitatively, can be mapped into a universal distribution if taking into account the statistical properties of each distribution. Differences in citation counts across disciplines are a well known bias, the roots of which lie in the different sizes of the fields or disciplines [14] as well as in different conceptual meaning of the citation itself [15]. In order to get rid of discipline dependent factors, the authors introduced a new Relative Indicator (RI) $c_f = c/c_0$ for each paper, where c is the number of citation the paper receives and c_0 is the average number of citations received by articles published in its field in the same year and writing a functional form for the distribution of RI as $F(c_f) = \frac{1}{\sigma c_f \sqrt{2\pi}} e^{-[log(c_f)-\mu]^2/2\sigma^2}$, where $\sigma^2 = -2\mu$ allows the expected value of c_f to be 1, thus allowing to compare the distributions across disciplines. Radicchi et al. also reported that the collapsing behaviour persists also when distribution from different years are compared, therefore suggesting that the functional form mentioned before is a *universal* curve, thus allowing to compare citation counts across fields and times in a fair way.

Field dependent patterns are also known to cause to disproportionate citation counts, even though they can be quantified and corrected for. This can be achieved by "imposing" a mapping between cumulative distributions of citations for papers published in a single category (i.e. subfields or fields) to the aggregated cumulative citation distribution [16]. For each field is therefore possible to assign to each citation count c' in the field cumulative distribution $P_f(\geq c')$ to the corresponding value c in the aggregated cumulative distribution ($P(\geq c)$) such that $P_f(\geq c') = P(\geq c)$. The relation between the two values for different fields is show in Fig. 2.1 as a quantile-quantile plot, in which it can be seen that the two citation measures are connected by a power law relation, therefore suggesting that the main difference between the citation distributions across fields lies only in a difference in each field's scaling factor.

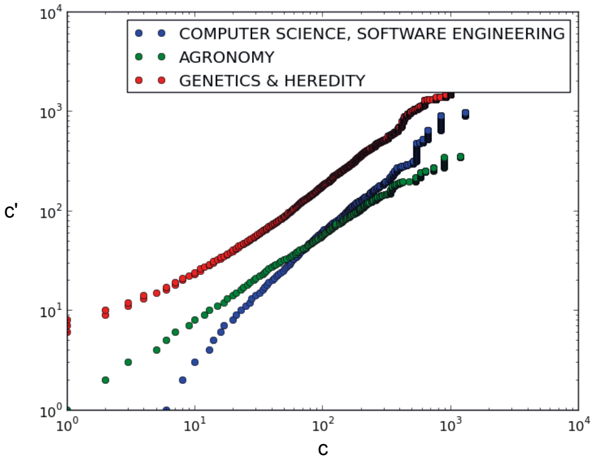

Figure 2.1. c' vs c adapted from [16] and reproduced with our data set. We can see that the scaling follows the relation: $c' = ac^{\alpha}$ where a is a pre-factor and α is a field dependent scaling factor.

2.2 Biases in citations

In 2005 Hajra et al. were [17] were among the first ones to suggest a temporal aspect in citation dynamics and decided to look at the impact that age has on citations. By looking at the citation dynamic of a set of papers, they found a critical time t_c of 10 years, after which the rate at which citations are gathered drop significantly, indicating that papers have approximately a *lifespan* of 10 years. In another paper in the following year [18], the authors suggest that the *rich get richer* mechanism might require to be connected with an aging of the publications in order to take into account the obsolesce of scientific publications. In Publication II we confirmed this property, showing that the typical life cycle of a paper is becoming shorter in time. Fig.2.2 shows the evolution of the time to reach the peak of citations for top papers in a selected number of fields.

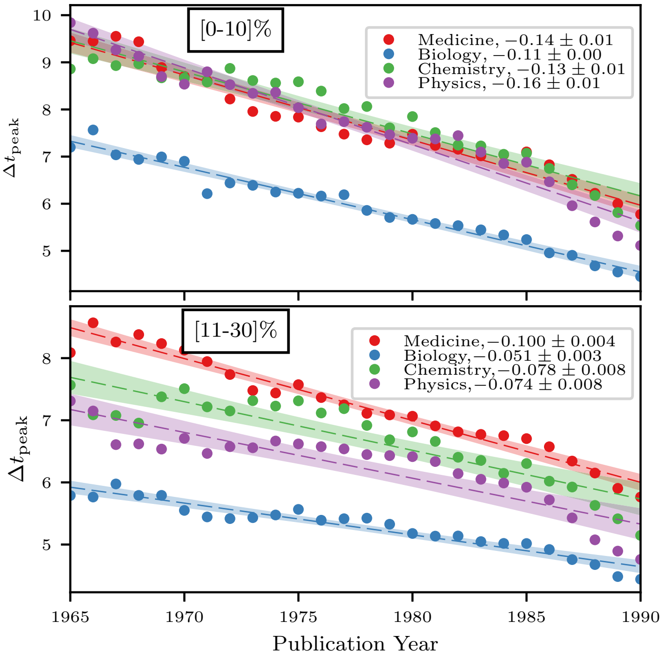

Figure 2.2. Time evolution of the mean values of time to peak Δt_{peak} for top 10% (top) and [11-30]% percentiles (bottom) of our ISI dataset. Δt_{peak} represents the time elapsed between the publication of a paper and the year in which it reached its maximum yearly citation count. The mean value $\langle \Delta t_{\text{peak}} \rangle$ decreases linearly in time. The linear fit, 95% confidence interval and the slopes of the linear fits are also shown. Figure adapted from Publication II.

While the average suggests that papers are being forgotten within a limited period of time, other works have been looking at the opposite phenomenon, the one of *sleeping beauties*, i.e. scientific papers that remained almost citationless for a long period of time only to become suddenly highly influential and cited [19]. The authors designed a Beauty coefficient defined as $B = \sum_{t=0}^{t_m} \frac{\frac{c_{t_m} - c_0}{t_m} * t + c_0 - c_t}{max\{1, c_t\}}$, where c_{t_m} is the maximum number of yearly citations gathered at time $t_m \in [0, T]$ and T is the time at which the coefficient is measured. The coefficient therefore quantifies how "unexpected" the citation history of a paper is, with $B = 1$ being the coefficient for a paper that grows linearly at a steady rate. One of the most interesting results of the study is that sleeping beauties, albeit appearing to be extreme cases, are impossible to distinguish from the core of all papers, as there is no minimum B^* value that allows to define a sleeping beauty as such. While most values of B are shown to be low, the authors conclude that it is an intrinsic property of scientific output to have a vast heterogeneity in the times at which recognition takes place. These results make particular sense for field such as Physics or Chemistry, where the theoretical and experimental sides of the same field are not always synchronized.

One of the most evident examples of this asynchronism is the recent experimental discovery of the Higgs boson, the existence of which was originally proposed in the 60s [20] but was confirmed only in 2012 thanks to the development of the LHC at CERN in Geneva [21]. The search of the boson was lagging so much behind that still 10 years after the theoretical breakthrough the hopes of a search for the Boson seemed remote despite phenomenological studies regarding its discovery had already started [22], as one of these studies points out [23] :

"We should perhaps finish our paper with an apology and a caution. We apologize to experimentalists for having no idea what is the mass of the Higgs boson, ..., and for not being sure of its couplings to other particles, except that they are probably all very small. For these reasons, we do not want to encourage big experimental searches for the Higgs boson, but we do feel that people doing experiments vulnerable to the Higgs boson should know how it may turn up."

The temporal aspect of recognition of older theoretical breakthroughs was a central source of inspiration for Publication I. In the paper we looked at the time lag between the publication of Nobel discoveries and the conferment of the prize, finding that it has been increasing at a very high rate, to the point where the original authors might pass away before seeing their discoveries empirically confirmed as shown in Fig.2.3. These findings led us to conjecture that we are potentially in presence of two opposite scenarios: either the frequency of groundbreaking discoveries is decreasing or, conversely, it could be that too many significant results are being published and that older discoveries are being awarded in order not to forget worthy winners.

Furthermore, one author might not be even aware of certain scientific works if he has not had the chance to read them or to search them efficiently. Even though the limitations of access to scientific knowledge might have become less relevant in modern times thanks to the rise of the Internet era and immediate access to online catalogues, at the same time the possibility to browse more recent material has consequently introduced a change in the way authors update their knowledge. The effects on the scientific community were rapid, as in 2003 already De Groote et al. [24] showed through a survey that general users of scientific material prefer digital copies to printed ones. The constant need for

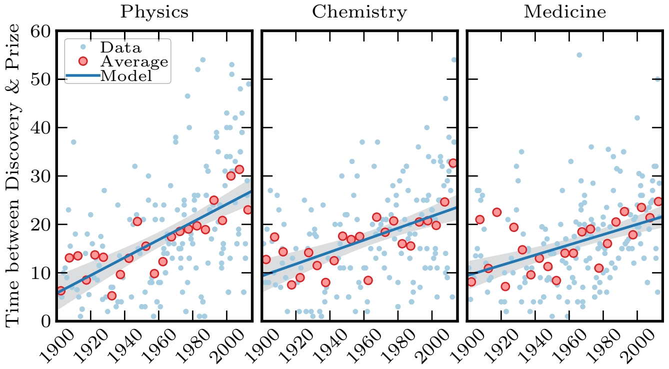

Figure 2.3. Time lag between discovery and Nobel prize vs year in which the prize was awarded for Physics, Chemistry and Medicine, created with data from Publication I. For each Nobel prize we searched bibliographic material on the author in order to identify one or more publications that could be directly associated to the awarding of the Nobel prize. The blue dots represent individual discoveries, while the red dots are a 5 year average over all awards in the bin. We can see a clear increase in average lag as well as the presence in more recent year of extremely high values (lag ≈ 50 years). Figure adapted from Publication I.

immediate access to recent scientific knowledge has become such a relevant aspect of science itself that it has led to suggesting a ranking of journals in terms of the speed at which their publications complete their cycle [25].

An interesting study in the impact of online available material on citation patterns came in 2008 when Evans [26] studied the effect of online availability of journal issues within the citation patterns of the journals and reported that the rise of online available publications shifted the citation patterns. The results showed that the more journals started to appear online, the more the reference list tended to be pointing at more recent discoveries and caused a *concentration* of citations towards fewer articles and fewer journals, an effect the authors claim is caused by hyperlinking, i.e. the search of further bibliographic material from the reference lists of papers previously read.

Recently however the claim has been challenged by Verstak et al. [27] as well as by Pan et al. [28]. Verstak et al. used Google Scholar Data to analyze all publications available between 1990 and 2013. The authors calculated the fraction of references in these papers pointing at least 10 years before the year of publication for each paper and found that such fraction is actually *increasing* in time. Furthermore, they noticed that the value of the change over the second half of the period studied was much larger than in the first, with the former matching the period in which digitalization has took place (2001-2013). The authors therefore concluded that the accessibility of older material has allowed scientists to cite the most suited paper that they were able to find, regardless of the time at which it was published. The latter paper by Pan et al. instead devised a model to test Evans' hypothesis which builds a citation network in which papers choose whom to cite both by "browsing" (i.e. by searching previous publications freely) and by a *redirection* link-formation mechanism in which knowledge is found by following the reference list of a source article previously browsed. By controlling the rate at which these two processes take place the authors simulated a spark in the redirection mechanism, representing the availability of online journals. The model showed that the redirection mechanism had very little impact on the average age of citations, while the growth of the system appeared to have a much more significant role.

The constant increase in scientific works might limit the ability to physically and mentally keep track of all relevant publications being published. This might be among one of the greatest limiting factors in citation patterns, as it has been reported [29] that scientists read more papers, yet dedicating less time on average to each one. The temporal dimension of the citation selection process has been the key source of inspiration for Publication II, where we suggest that the increasing number of publications causes a constant shift in focus towards more recent papers, therefore shortening the citation life cycle of papers both in terms of time to reach their peak in popularity, as well as in terms of time needed to stop gathering significant citations after the peak. Fig. 2.4 shows the main results of the analysis.

Another aspect that influences citation choices is one that looks at the role that the individual authors play. Science is not only a philosophical endeavour, but also a social system where scientists personally interact and collaborate and therefore are more exposed to works coming from a familiar set of collaborators or, in general, people working in the same area of research. Early research in fact showed that [30] intellectual ties based on shared content surpass friendship as a predictor of reciprocal citation. Similarly, Persson et al. [31] showed in 2004 that collaboration leads to a positive effects in the success of a paper, in particular if the authors come from different countries. This can be seen as a success linked to the possibility of the same work to be pushed forward at twice (or more) the same rate as a single author paper in different "market pools" of customers, i.e. potential citers. Furthermore, in 2004 Glänzel et al. reported that multi-authorship increases the chances of self citation [32], with the number of authors not being a factor though. However, the authors point out that the most dominating contribution of multi authorship is the increase in foreign citations, thus showing the social contribution of a multi author paper in terms of geographical advantage.

The topic of self-citations is a highly debated one in a world where citation metrics are used as tools to quantify careers and quality of research. The same author showed in another paper in the same year that self citations are an *"Essential part of scientific communication"* [33], but that its contribution plays a higher role in the immediate times after publications. This result, linked with empirical evidence of self-citation being correlated with publishing on average in journals with relatively low impact shows that this trend might be linked to the need of a "push" in fame, hoping for success to accumulate from there. However, while self-citation does appear to have an impact on citation counts, it is not clear whether the correlation is linked to a matter of *visibility*, i.e.

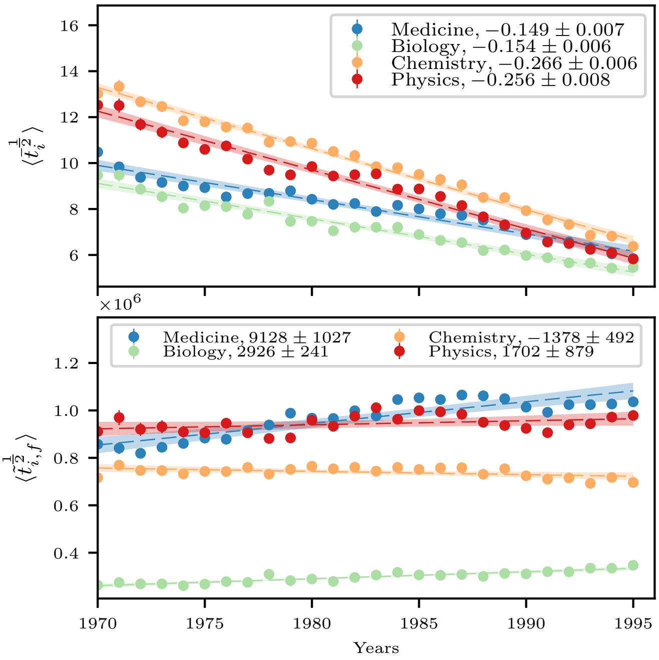

Figure 2.4. The evolution of the half life of papers after the peak $\langle t_i^{\frac{1}{2}} \rangle$ in terms of absolute time (top) and $\langle \tilde{t}_{i,f}^{\frac{1}{2}} \rangle$ in terms of the number of publications (bottom) for the four different fields and for the top 10% percentile. For each paper we calculated the time required for the publication to drop below half of the number of citations gathered in its peak year. We then proceeded to average the values for papers published in the same field and peaking in the same year. The half life has been calculated both in terms of number of years and in terms of number of paper published within the field in the same time interval. The linear fit, 95% confidence interval and the slopes of the linear fits are also shown. The dashed line represents the linear fit. Despite its noisy behavior, the renormalized half-life shows a relatively stable trend throughout the years, possibly with the only exception of Medicine and Biology, which show a slightly rising pattern for recent time. Figure adapted from Publication II.

trying to put forward one's results as a "bandwagon" effect, or rather a matter of *quality*, as one author mentions its own best works as a basis for future ones [34]. More recent results confirm [35] that the trend is still significant, yet retaining different patterns in different fields, due to the possibility of certain fields to have many groups working on independent topics, thus focusing the selection of cited material from a smaller subset of works. The authors also report that a higher propensity in inter-author-citations leads to a higher chance of inter-citations of the second order, with collaborators of collaborators being more likely to be cited.

Authors might also influence their own career retroactively as shown by Mazloumian et al. [36]. The authors found that groundbreaking results by an author have a positive impact on their own previous literature, therefore creating a status of authority for the author even though the earlier works might not be necessarily related to the successful recent ones both in terms of topic and intrinsic scientific quality. The role of prestige in science is so critical that it has been suggested to also be a bias within the peer review mechanism [37]. This psychosociological mechanism that enhances the career of already successful scientists based on their academic reputation is often called *Matthew Effect* and its impact on science has been discussed since the 60's [38]. In general, a citation bias towards successful papers (preferential attachment) and one towards successful authors (Matthew Effect) shows that the citation mechanisms are not only based on scientific necessity, but are also based on individual and collective aspects that emerge from the human interaction between scientists. Finally, it is worth to mention that there are plenty of other factors that influence citations, such as journal-dependent factors, field-dependent factors and technical ones [39], which will not be analyzed for the sake of brevity.

2.3 Modeling

The previous section showed how many factors and biases play a role in the mechanisms underlying the decision of which papers will appear on a reference list, with empirical results showing heterogeneous results within the same field of analysis. It is therefore not surprising that the pursuit for a mathematical model that could correctly reproduce the properties of citation mechanism has been a challenging one, which scientists however were eager to undertake in order to shed more lights on the way science itself works, focusing in particular on the temporal aspect of the models.

The earliest and most successful attempts at modeling citation dynamics lie in the *rich get richer* or, technically speaking, *preferential attachment* mentioned in the previous sections . Despite the original idea was already formulated in de Solla's work [4], it was Barabási in 1999 who was able to mathematically describe it exhaustively [11]. In his work, Barabasi suggests a model (PAM) in which the probability (or attachment rate) A of a paper of receiving another citation from a new paper is directly proportional to the number of citations c citations previously collected: $A(c) \propto c$. This mechanism is able to explain the citation distribution both from a qualitative point of view (its fat tailed behaviour) as well as numerically, confirming an expected value extremely close to 3 for α. Interestingly, the model was applied to a vast amount of complex systems, with particular success in biology [40, 41], of which citation dynamics represent one of the examples.

A confirmation of the validity of the preferential attachment mechanism came in 2005 with Redner [12], who reported that the attachment rate is indeed linear, leading to a double paradox: the linear attachment rate shown by the data should lead to a power law distribution for citations, while data shows that the form is log-normal, which in turn would require an attachment rate of the form $A_c = \frac{c}{1+a\ln(c)}$ with $a > 0$. Despite confirming empirically the validity of a linear form of preferential attachment, Redner suggests that the underlying assumptions behind the preferential attachment model, when applied to science, might be not completely realistic, as the model implies a full knowledge of all the corpus of existing papers, a challenge which has its limitations both in terms of accessibility as well as in terms of memory.

As we saw in the previous section however, it is fundamental to introduce the question of time dependence within the modeling framework. While theoretical works tried to tackle the topic from a purely mathematical standpoint [42, 43], it was Hajra et al. [17] in 2004 who applied it with success to the modeling of citations. The authors followed the previous theoretical works and formulated a functional form for the attachment rate of $\Pi(c,t) = C(c)T(t)$, where $C(c)$ and $T(t)$ are generic functions and where the attachment rate is assumed to be separable. The authors then tried to identify the functional form for the temporal aspect that would best fit the data through the analysis of the distribution of citation ages $Q(t)$, i.e. the raw distribution of the fraction of citations with age t. In order to do so, the authors took into consideration the stochastic nature of the rate at which new citations appear, i.e. the rate at which new papers are published. Therefore by empirically estimating from their data sets a publication rate of $n(t) = a(1-e^{-bt})$ they were able to renormalize the distribution and obtain a functional form of $T(t) = \frac{Q(t)}{n(t)}$. Comparing the model with the collected data, the authors identified two distinct regimes of power-law decay of the distribution: $T(t) \sim t^{-\alpha_1}$ for $0 < t < tc$ and $T(t) \sim t^{-\alpha_2}$ for $t > t_c$ where $t_c \sim 10$ is the expected lifespan of a paper mentioned earlier.

In Publication II we proposed a model for the process of gathering new citations as a *counting process*. In this ultradiffusive framework, the arrival of a new citation is hypothesized to be correlated to an earlier event or a combination of events. Therefore, ultradiffusion proposes that the pattern of events emerges as a consequence of an underlying hierarchy of states, in which a more recent event is more likely to affect the future ones. Our results, that show an exponential fall in citation after reaching the peak, which is slowly transitioning into a power law pattern is coherent with the hypothesis of an ultradiffusive process driving the attraction of new citations. This framework is known to be able to explain the evolution of the response to new pieces of information online [44], allowing us to draw a comparison between the way in which attention is dedicated to new publications and the way readers react to news.

A further improvement on the PAM came in 2008 with a work by Wang et al. [45]. Their model proposes to not separate globally the dependence of the attachment rate on the two variables, considering the aging process to be related not to the whole paper, but to the citations themselves. The logic behind this idea is that a paper that has received a lot of attention lately (a sleeping beauty for example) will be more likely to gather new citations if compared to a paper published in the same year, with a similar citation count, but having failed to receive citations recently. Therefore, the authors express the rate as $\Pi(c,t) \propto \sum_t c_i f(t_i) \propto \sum_t c_i exp(-\lambda t_i)$, where k_i are the citations gathered in year t_i and the exponential form for the weights is taken from fitting data, a scheme they call Gradually-vanishing Memory Preferential Attachment Mechanism (GMPAM). While the empirical data shows a good accordance the model, the authors admit that the model is somewhat excessively complicated, as it requires to calculate weights for decades of citation data coming from different citation pools (field and geographical biases above all) that require to fine tune the value of λ case by case. The authors therefore proceed to simplify the model, by observing that the most significant temporal contribution to the attachment rate comes from the most recent number of citations, i.e. the number of citations gathered in the last year. The temporal aspect therefore it's taken to be as a *memory effect*, that makes the older citations be "forgotten", giving priority to papers that are riding a popularity wave. The updated model, called Short-term Memory Preferential Attachment Mechanism (SMPAM) thus expresses the attachment rate as $\Pi(c,t) \propto c_{t-1}$.

Similarly, other authors have decided to focus the modeling part only to reproduce certain aspects of the citation dynamics with still a focus on the temporal aspect. In 2001 Burrel was able to confirm that a stochastic process that assigns citations to publications based a non-homogeneous Poisson process [46] is bound to produce articles that will remain uncited. In 2009 Wallace et al [47] tried to model the citation distribution of publications by separating the citation curve in different areas, developping in particular a model able to quantify the impact of uncited papers in the citation distribution. The authors hypothesized that the probability for a certain paper to receive an initial citation depends only on the number of articles N_A published in the same year and the number of references N_R available in the following year, with citations being given randomly through a Poissonian distribution, given the size of the two variables. The authors then limit the probability of citing an uncited paper to the field-dependent rate at which uncited papers are cited for the first time. It therefore follows that the pool of available references is reduced to $\beta_I N_R$, where $\beta_I \in [0,1]$ is extracted from the data, and that the probability for a single paper to fail to receive any citations is: $\Phi_I = e^{-\beta_I(N_R/N_A)}$.

In 2009, Newman [48] published a study which added to the temporal aging process the aspect of novelty, the so called *first mover effect*. The idea behind the work is that science is based on the production of new results and therefore there is an intrinsic advantage in the being the first ones to publish a new result in a field, since future works are bound to cite the paper introducing the novelty. In his paper, the author works with previous models based on preferential attachment to build a new one where on average newly published papers cite m earlier papers, chosen proportionally to the number of citations k they already have, plus a variable r needed to ensure that uncited papers still have a nonzero probability of being cited. From this model one can calculate the average number of citations γ a paper is expect to receive at time t as: $\gamma(t) = r(t^{-1/(\alpha-1)} - 1)$, where $\alpha = 2 + r/m$. Therefore, it follows that older papers (i.e. $t \to 0$) should on average receive far more citations than those published later, even taking into consideration the the fact that later papers have less time to gain citations.

These results are somewhat in contrast with the previous discussion regarding obsolescence and the time span of papers. However, Newman himself points out that the first mover advantage is limited to scenarios in which the results are not part of a larger, already established field, but rather represent the emergence of new subfields or fields altogether, as their analysis of citation data in fact seems to confirm.

In 2011 Eom and Fortunato [49] published a paper in which the aspect of the *burstiness* in science is tackled. Burstiness is a sudden and intermittent modification of the frequency of an event, which has been known to play a fundamental role in many human dynamics [50, 51]. In this context, burstiness represents all sorts of inhomogeneous fluctuations that lead to a sudden and unexpected rise in the citation count of a paper, which can be expressed as $\Delta c/c = [c(t+\delta t)^i_{in} - c(t)^i_{in}]/c(t)^i_{in}]$, where $c(t)^i_{in}$ is the number of incoming citations a paper received at time t measured in years. This rate therefore measures the relative change in citations during the period of time δt, compared to the history of citations of the paper. Data shows that the distribution of these rates is fat tailed for $\delta t = 1$, showing therefore that it is possible for a paper to suddenly receive orders of magnitude of citations more than they ever did, especially during its early years. Similarly to what happens to sleeping beauties, burstiness shows that there can be stochastic driving forces that cannot be ignored and that a linear model with no memory or time dependence cannot grasp. The authors therefore propose a model still based on the preferential attachment model, where however each papers has an intrinsic *attractiveness* that depends on time. The result is a model in which a new paper i cites m new papers, with the probability of a certain paper j to be cited described as : $\Pi(i \to j, t) \propto [c^j + A_j(t)]$. For the form of the attractiveness the authors assume an exponential decay $A(t) = A_0 exp^{-(t-t_0)/\tau}$, where τ is the time scale at which the temporal dimension plays a role, with initial attractiveness taken from a power law in order to best fit the data. Once again, we have a model where a linear preferential attachment is mixed with a temporal dimension, which in this case takes into account random fluctuations of the citation history of the paper that alter the expected individual citation trajectories. Attractiveness can be seen as proxy of an intrinsic *quality* of the paper, which is explicitly separated by the success of a paper in terms of citation. The model therefore suggests that citations do not represent the absolute measure of the quality of the paper, but that rather they are a probable (but not guaranteed) consequence of papers of high quality (attractiveness). However, with citations and preferential attachment still being a fundamental driving force of the citation market, an initial failure to gather an initial minimum number of citations might be sufficient to prevent a high quality paper from rising to notoriety.

In 2015 Wang et al. [52], including the original proponent of the Preferential Attachment Model Barabási tried to further expand the concept of separating the driving force of citation and the one of fitness of the individual paper, by proposing an attachment rate of the form: $\Phi_i(t) \propto \eta_i P_i(t,\mu_i,\sigma_i)c_i$, where η is the fitness of the individual paper and $P_i(t,\mu_i,\sigma_i)$ represents the aging process of the ideas introduced by the paper. The separation of fitness from aging (i.e. it's not the fitness that decays, but rather the *novelty*) comes at a cost, as the authors needed to introduce two new parameters, represented by the immediacy η of a paper and its longevity σ which determine the time at which a paper reaches its peak of notoriety and how long its notoriety will last respectively. The model is therefore able to predict the future citation trajectory of a paper, given a previous window of time during which its intrinsic parameters can somehow reveal themselves and be quantified through a least square fit method. Furthermore, the authors managed to quantify the importance of the individual contributions within the attachment rate formula, finding that the dependence on the number of citations (i.e. the classical model) is triggered only when a paper crosses the threshold of seven citations, below which it's the paper attractiveness that dominates.

3. Network Structure of Science

De Solla's seminal paper [4] begins like this: *"This article is an attempt to describe in the broadest outline the nature of the total world network of scientific papers. We shall try to picture the network which is obtained by linking each published paper to other papers directly associated with it."*. Already at the beginning of the study of scientometrics it appeared evident that science needed to be tackled from a global perspective, analyzing the connections that link scientific papers to one another. Similarly, two co-authors of the same paper can be linked together, as well as two scientists who have collaborated with the same scientist as the famous Erdős number grasps [53] [1]. In general, the intrinsic collaborative nature of science either by cumulative contribution (the shoulders of giants) or by direct collaboration has led to the creation of a massive scientific network that can be analyzed in many of its levels, where both its nodes and links can take many forms, with nodes representing papers as well as authors, institutions or countries and links representing citations, co-authorship, shared funding etc.

Graph theory showed for the first time the potential of network research for practical problems in the famous work by Euler in 1796; by simplifying the bridge and road structure of the city of Königsberg in terms of nodes (land masses) and links (bridges), the Swiss mathematician was able to negatively answer the question: is it possible to perform a path around the city that crosses each bridge of the city exactly once? For a long time graph (or network) theory remained confined mainly as a branch of topology in theoretical mathematics [54] until the middle of the 19th century when the earliest structured books appeared [55, 56], allowing the developments in the theory to spread to new fields [57], including sociology, where researchers understood that a matrix based representation, i.e. one of the underlying bedrocks of network theory, of social ties could be beneficial for the study of social structures [58, 59]. The breakthrough came in 1959 with Erdős and Rényi's work on random graphs

[1] The Erdős number measures the distance in terms of collaborative steps between the Hungarian mathematician Erdős and his direct or indirect collaborators. Anyone who has collaborated with him has a Erdős number equal to 1. All their collaborators have a EN of 2 and so on.

[60] in which the authors studied the invariant properties of graphs generated through a stochastic model that distributes a fixed number of links across all possible node pairs. The ER model turned out to have strong analogies with statistical mechanics [61] and was later used as a fundamental tool for studies that required a network based structure, in particular for models in epidemiology [62, 63].

In general, the ER model allowed the rise of what are called *generative models*. These models aim at reproducing the statistical properties of the observed networks [64], yet keeping the most important features (usually the degree distribution or the average degree) of the network statistically constant, while allowing for the edges to be distributed at random. Generative models therefore act as tools for generating null-hypothesis that can be tested statistically, allowing to identify which properties in real networks are statistically relevant, with applications to multiple fields [65, 66]. Among the attempts, de Solla Price contributed with the earliest definition of the rich get richer mechanism [10] that would later be made popular by Barabási and Albert, who showed its potential [11] as a tool to describe the emergence of scale-free networks. Barabási and Albert's paper was part of a period of extreme interest for network theory studies as the rapid accumulation of data of large networks thanks to the digitalization of society, allowed for the first time to provide a robust set of data that could be used to test previous models. While the ER model had been extremely successful due to its simplicity, the evidence of different properties in real networks required the development of new models, which rapidly took place [67, 68]

Since then, network theory has been applied in a large spectrum of fields, dealing with non-trivial network structures that required methods and algorithms tailored to specific types of network problems, leading to a whole new field, often referred to as *complex networks*, in order to differentiate it from Graph Theory. As the theory developed, the application of its methods to publication data became a fertile branch of the field. This Chapter will first go through the basics of network theory, in order to provide a mathematical foundation for the rest of the chapter, in which the most significant applications to scientific networks will be discussed.

3.1 Networks

A network, also called graph, is a collection of nodes connected by links. Mathematically it is represented by $G = (V,E)$ where V is a set of N nodes and E is a set of M links (or edges) connecting pairs of nodes. A convenient way to represent a network is through its *adjacency matrix* A, which fully describes the graph. Its elements a_{ij} are 1 if there is a link connecting node i and node j and 0 otherwise. If A is symmetric the graph is undirected as all of its links go in both directions. It is often assumed that there are no self loops, i.e. $a_{ii} = 0$ for all i. In this simplest scenario, the elements of the matrix are usually binary and symmetric, thus only indicating whether two nodes have a connection or not. However, more sophisticated networks can be built by modifying these conditions: *Directed graphs* take into account the directionality of the links by dropping the symmetry requirement, while *weighted* graphs drop the binary requirement for the elements of the matrix, thus quantifying the "strength" of the link. An example are mobile call networks, in which a_{ij} can indicate the number of calls between user i and j, or the total time spent between two users [69]. Networks in which most elements of the adjacency matrix are 0s are usually called *sparse*, while in the opposite case they are called *dense*. Sparse matrices, which are not rare at all [70], can represent a problem computationally in terms of storage space since, if stored in matrix form, N^2 entries need to be stored, most of which do not carry information. Fortunately, the disadvantage can be turned in an advantage by using *adjacency lists* in which each row i enumerates the neighbors of the node along with the value of the edge in case it is required. Recently, there has been a need to analyze many different kinds of network structures. For example, temporal networks take into consideration the intermittent activity of the edges in the network, thus adding a temporal dimension to the analysis of complex networks [71]. Multilayer networks instead deal with systems in which nodes exist in one or more of multiple layers and where links can connect nodes also across layers [72, 73]. Such networks can be useful to analyze interactions in social systems, where each layer represents a different kind of interaction and where not all users are equally active in each layer, or might not be active at all in some of them [74].

3.1.1 Degree

The degree k_i of a node is the number of nodes that node i is connected to. It can be derived using the adjacency matrix A as $k_i = \sum_j a_{ij}$, i.e. the sum of the nonzero elements of row i. In case of a directed network two separate degrees are considered : k_i^{in} and k_i^{out}, which differentiate between the degree calculated respectively over the columns or the rows. The average degree $\bar{k}$ of a network is the average value of individual degrees $\bar{k} = \frac{\sum_i k_i}{N}$, where N is the number of nodes in the network. Again, it is possible to define an average $\bar{k}_i^{in}$ and $\bar{k}_i^{out}$ for directed networks.

When analyzing a large network, it can be useful to look at the overall distribution of the degree values for the nodes of the network, as with an increasing number of nodes it becomes necessary to analyze them statistically. In the ER model [2] each link exists with probability $\frac{M}{\binom{N}{2}}$, leading to the probability of node i to have degree k to be the probability of having k times successful Bernoulli trials, thus converging to a Poissonian distributions as the size of the network grows, with $\bar{k}$ remaining constant. However, empirical evidence [76] has shown that real world networks have a dramatically different behaviour when it comes

[2] This formulation was presented in the same year by Gilbert [75] and is statistically equivalent to the ER model.

to degree distribution.

While the ER model predicts a large amount of nodes sharing similar degree values, social, biological and transportation network among others, revealed themselves to have fat-tailed distributions [77], i.e. they showed the existence of nodes with large degree called *hubs*, along with a vast amount of nodes with low degree values. In 1999, Barabási and Albert proposed a different model, in which the network is generated by adding new nodes and connecting them proportionally to the degree of the previously existing nodes, through the Preferential Attachment Method already introduced in the previous chapter. In Fig.3.1 we can see a comparison between the appearance and the degree distribution of a random networks compared to a scale-free network.

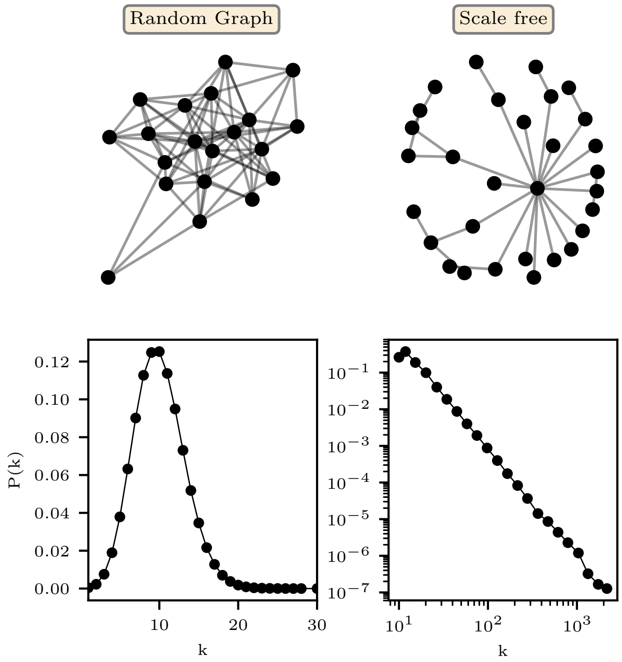

Figure 3.1. Difference in topology and degree distribution between a random graph (left) and a scale-free network (right). The random network has its degree distribution heavily centered around its average, with no significant outliers. In the scale-free model instead, degrees can span multiple orders of magnitude.

Another fundamental property of degree is linked to the concepts of *assortativity* and *resilience*. Assortativity is used to investigate what is the tendency in a network for nodes with similar degree to be connected [78, 79] and is therefore often expressed as degree-degree correlation. In a network with high assortativity, high-degree nodes are likely to be connected and tend to avoid connections to low-degree nodes. Similarly, a network is disassortative if high degree nodes tend to avoid being linked to each other and prefer being connected to lower degree nodes. In both the ER and Preferential Attachment models, there is no correlation between degrees; in the ER model links are given randomly, thus an absence of correlation is to be expected for large graphs, while in the PA model the evidence is less trivial, but it comes from the fact that hubs have a tendency to get links from all new nodes, thus failing to select connections to specific nodes. Interestingly, real life networks show different scenario, with certain networks being assortative (power grids, social networks) and other disassortative (WWW, protein-interaction networks), thus requiring more sophisticated models to be able to reproduce these features [80]. A direct consequence of assortativity is resilience, i.e. the ability of a network to resist the attack or failure of random nodes. In a air transportation network for example, this corresponds at how the passenger traffic is affected by the closure of randomly selected airports. Numerical simulations [78] show that a high assortativity is linked to a better chance to resist attacks due to the fact that hubs, which are often fundamental as they allow to distribute "services" to the periphery of the network, are likely to be connected to each other, thus creating dense cores of highly connected nodes that keep the structure of the network efficient. In disassortative networks instead, hubs are fundamental local service providers and, if shut down, are more likely to cause an interruption in services. Unfortunately, many communication networks are disassortative [81] and have therefore been often subject of systematic failures [82] due to their structural inefficiency.

3.1.2 Clustering, paths and distances

The clustering coefficient measures how likely two nodes within the neighbourhood of a node are also be connected [67]. Let's consider a node with k neighbours. Among these neighbours there are $\frac{k(k-1)}{2}$ possible links, i.e. the number of ways 2 nodes can be selected if there are k nodes, out of which only E_i are present in the network. The CC is defined as the ration between the two terms:

$$C_i = \frac{E_i}{\frac{k_i(k_i-1)}{2}} \tag{3.1}$$

In case of weighted and directed graphs the concept can be generalized in multiple ways [83]. The average clustering coefficient of a network is the average $C = \sum_i C_i/N$ of the individual clustering coefficients. The global clustering coefficient is a similar measure as the average clustering coefficient which looks at the clustering of a network from a geometric point of view. It is defined as the fraction of triplets (i.e. a set of 3 connected nodes) that actually form a triangle and can be applied to both undirected and directed networks [84]. In an undirected network the average path length between two nodes is defined as $l = \frac{1}{\frac{n(n+1)}{2}} \sum_{i \geq j} d_{ij}$, where d_{ij} is the length of the shortest path between two nodes. In case the graph is not connected (i.e. there are parts of the networks that are separated), the value of the average path length diverges and is therefore convenient to compute it individually for each subgraph of the network. The diameter, D, of a network is defined as the maximum shortest path between any two nodes in the network. Its name recalls the topologic properties of circles as it represents the approximate linear size of the network.

In 1998 Watts and Strogatz published a paper that showed how the currently available models based either on regular lattices or on random graphs were unable to grasp the properties of real networks in terms of clustering coefficient and path length[67]. While their analysis of diverse networks (power grids, biological networks, film actors) showed large CC and short paths, the ER model [68] is bound to generate networks with average path length $\propto log(N)$ and have an extremely low value for the CC. They called their networks *small-world networks* in reference to the famous social experiment of the six degrees of separation [85], which was the first attempt at calculating path lengths in social networks. They proposed a stylized model based on a regular lattice, thus guaranteeing high clustering, with a random rewiring of each link controlled by a parameter p. The value of p therefore allows the transition from a regular lattice ($p = 0$) to a random network $p = 1$. As p increases from 0, local clustering remains high while paths between distant nodes cause a significant reduction of the average path lengths. With this simple model Watts and Strogatz managed to show how even a small number of short cuts can transform a sparse, locally clustered network in a small-world one.

3.1.3 Communities and modularity

Between 1970 and 1972 Wayne W. Zachary collected data about the interaction between 34 members of a karate club, during which two instructors had an argument, leading to a split of the group into two, with half of the group remaining in the club with one instructor and the other half leaving it [86]. Based on the difference between the interaction patterns, Zachary was able to devise an algorithm able to automatically detect in which half a node would lie. This became the first example, and later the benchmark, of a *community detection* algorithm [87]. The idea behind community detection is that networks can be organized in locally highly connected clusters separated one from the other, known as communities. Real world examples are abundant: metabolic networks are organized into small, highly connected modules [88], urban areas and societies can be structured in large groups divided by language [89], and also network scientists are organized in communities [90]. While communities are easy to qualitatively define, their mathematical definition has been the source of debates as, like in the Karate Network splitting in two roughly equivalent groups, one needs to possess previous information in order to know how many communities are to be found and what their typical size is.

As new algorithms attempted to find the most optimal division of the network in communities, it became therefore necessary to develop a method able to grasp the quality of the partition of the network. Among the various methods, the most popular one is the one of *modularity optimization* [87]. This method, introduced by Girvan and Newman in 2002 [91] is based on the idea that a good partitioning

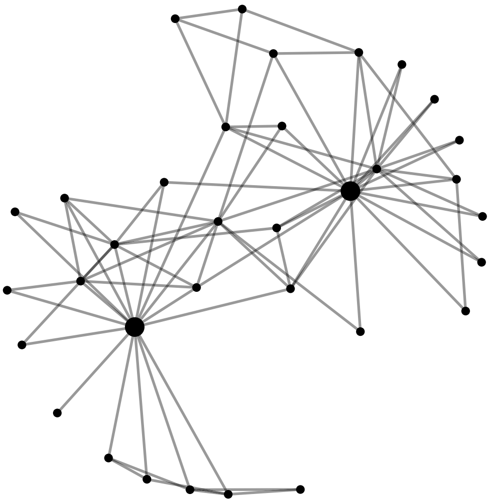

Figure 3.2. Visualization of the karate network based on the data from [86]. The network is visibly structured around the two hubs (larger nodes), with clustered communities around each hub and a few nodes acting as intermediaries between the two communities.

maximizes the amount of edges within a community and minimizes the amount of links towards the outside of the community. Modularity is therefore calculated as the difference in number of edges within a cluster and the expected number of edges that one would found in a similar network in which individual nodes retain their degree, but the edges are randomly rewired. In Publication III a similar idea was used to investigate how dense the subgraph of Ego Networks, the graph formed by the neighbours of a specific individual (the ego) and by their mutual relationships, is. The EN is the realistic, local perspective of a given node representing the information that it might use in basic decision processes. We calculated for new nodes joining the EN the fraction of references that stay within the EN, thus quantifying how modular the EN is and how its modularity evolves in time. We showed that the EN has a sharp initial growth in modularity that saturates within 10 years, before gradually decreasing as shown in Fig.3.3.

Unfortunately, despite its simplicity, modularity also offers some limitations. Fortunato et al. showed in 2007 that modularity optimization is bound to have a resolution limit, i.e. a minimum size of communities under which the method fails to detect communities [92], which can represent an issue as real networks can be organized in hierarchical or tree-like structures [93]. Furthermore, such resolution limit depends on the size of the network; as a network increases in size the null model might expect two clusters to have a very low probability to be connected, therefore allowing one single connection between them to be seen as a strong statistical indicator of modularity, thus merging the two clusters. Even by trying to introduce a resolution parameter in order to find clusters of various sizes, problems such as merging of subgraphs and splitting of graphs arise [94]. Furthermore, another key limitation is the presence of multiple suboptimal solutions [95] that still offer good results. While other methods are being introduced with good results, they all come at a cost somewhere, due to the intrinsic loose definition of a community, thus forcing the scientists to perform a trial and error analysis based on the cumulative information gathered in the process [96]. To summarize, there is no "Free Lunch" in community detection [97].

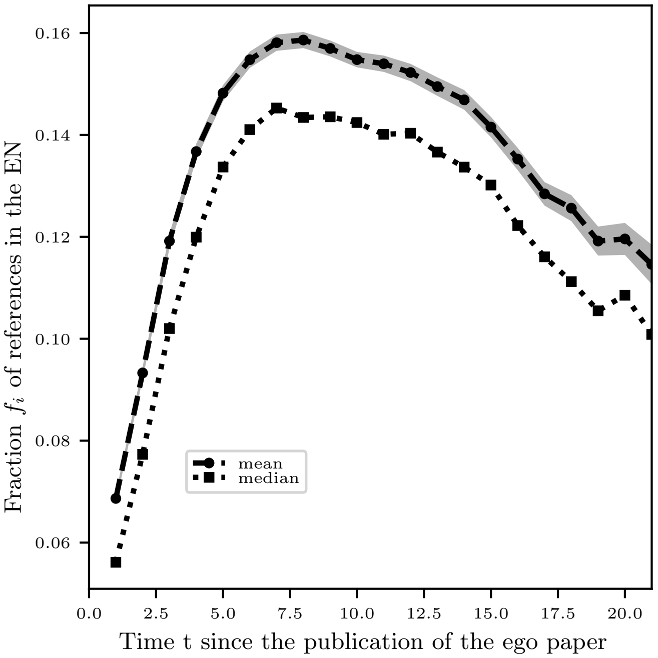

Figure 3.3. Time evolution of the mean and median of the fraction f_i of references of papers of the full Ego Network belonging to the Ego Network as a function of the number of years since publication. In this framework, f_i is the EN equivalent of the modularity of the community that is formed around the original paper. In the first years f_i increases significantly, peaking after ≈ 7 years, after which a constant decrease takes place. Interestingly however, the EN is also getting bigger in size, thus potentially allowing for more references to be part of the EN. Figure adapted from Publication III.

3.2 Author networks

As we have seen, network theory provides a solid framework with which to investigate social structures. It followed therefore that scientists could use the very same methods to investigate the social structure of science itself. The main candidate for such analysis are author based networks, i.e. networks in which nodes are represented by individual scientists that are connected according to similarity in their publications.

The most straight forward approach is the one to considers co-authorship networks, in which links are assigned between scientists who collaborate in the writing of a single paper. The first study in the field was performed by Newman in 2001, by studying a dataset of over 2 milion papers and 1 milion authors in Physics, Computer Science and Biomedical research [98]. This work allowed for the first time to quantify the collaborative structure of science with the newly formulated tools of network science. The data showed that the degree distribution, i.e. the number of collaborators for a single author, follows a power-law behaviour with an exponential cutoff, a result coherent with a power-law degree distribution, with the cutoff being due to a size restraint in the system. The author also reports that the network of scientific collaborations shows a small-world structure, with authors being no more than five or six steps apart from each other. The network showed also an interesting tendency for authors to cluster, even though this might be biased by the presence of papers written by 3 or more authors, which, by the network construction rules, create triangles in the network. Newman's work showed the intrinsic social nature of science as a network of collaborating nodes, with a structure that is coherent with a PAM in which authors with most collaborations are more likely to collaborate with new scientists. However, from a theoretical point of view, it fails to find an explanation for the coexistence of a power-law degree distribution and the intrinsic community-based structure, a feature absent in the PAM.

The matter was further analyzed by Barabási et al. [99], who confirmed the clustering nature of co-authorship networks with a caveat: clustering, as well as other key properties of the network, are time dependent, therefore providing only partial information about the true structure of the network. This work, while reinforcing a preferential-attachment approach to the evolution of co-authorship networks, once again introduces the matter of time in the exploration of properties of the scientific community.

It has been suggested that a major role in the temporal aspect of co-authorship networks may reside in the evolution of the individual careers of the different authors [100]. Sociological considerations [101] can support the hypothesis that the preferential attachment method, that is the phenomenon by which authors with many collaborations are more likely to have new ones, is the the driving force only of collaboration only for scientists in the middle of the career (thus also in the middle of the distribution). The tails of the distribution instead are dominated by either established scientists, who don't require to build up their network anymore, or newcomers who instead fail to act as attractors in the network. It therefore follows that one cannot investigate the social structure of science in snapshots, but rather needs to follow its temporal evolution as *"networks change over time, both because people enter and leave the professions they represent and because practices of scientific collaboration and publishing change"* [102].

Furthermore, one needs to step at a deeper structural level: while co-authorships provide the basic framework, it is important to differentiate between the various substructures that exist within a network as evidence shows that the local structure of the network has an impact on the citation and co-authorship patterns [103]. In fact, co-authorship practices are extremely heterogeneous across fields, as in certain applied sciences it is not rare to find papers co-authored by tens of authors, thus putting into question the ability of this approach to reflect the social structure of science. In fact, networks of different size need different collaborative behaviours for the their community structure to persist in time. While smaller collaborative groups tend to be based on a core of strong relationships that are self-sufficient, larger groups need a more dynamic structure that reaches out to new members in order to survive, similarly to what happens in mobile communication networks [104].

Even though the co-authorship network is purely abstract in its formulation, it is possible to merge it with physical data, e.g. the location of the institution in which the authors work, allowing to add a geographic dimension to the analysis. Relocation is common in academia, even though scientists usually are not likely to cover long distances, and can play a crucial role in one's career [105]. Similarly, the choices of collaborators are also affected by geographical considerations that can be linked to policy making from individual countries or unions [106, 107].

3.2.1 Ties and careers

In a framework in which the career and the connections of individuals change structurally over time, it becomes therefore fundamental to investigate the different nature of the links that connect different authors at different stages of their careers; after all science is not only driven by purely intellectual but also by more practical driving forces, such as economical and political matters that can also alter the paths of individual careers [108, 109], thus affecting the structure of collaborations both locally and in time. Similarly, as the network structures are known to influence team-performance [110, 111], it is natural to conjecture that these kinds of mechanisms are reflected in the data of scientific collaborations.

In order to better understand such effects it is beneficial to investigate the role of the *strength* of the ties between authors as a measure to identify which connections are more productive and represent a stronger tie within the sphere of scientific collaboration. This can be done by building a weighted network, where the weight of each link is defined as $w_{ij} = \sum_p \frac{1}{n_p - 1}$ where p is the set of papers where authors i and j collaborate and n_p is the number of co-authors of paper p. Contrary to previous results in social networks [69], collaborative networks show a unique characteristic: weak ties form the core structure of dense neighbourhoods, with strong ties connecting different neighourhoods. This effect is considered to reflect the hierarchical and temporal dimension of scientific careers: as senior researches build strong ties with each other over time, they form research groups composed of young researchers [112, 113]. Even though it is only a few strong links between senior scientists that keeps the scientific network of authors together, simulations show that they are fundamental for the efficient spreading of information through the network.

In an academic world where most junior scientists drop out [112], which is hierarchically and sometimes unequally structured in its hiring system [114] and in which early developments can lead to a cumulative advantage in a career [38, 115] it appears evident that the evolution of the social and collaborative structure of scientific interaction is closely related to the evolution of the individual careers of the prominent scientists: their moving forward in the hierarchy of science, projects their connections to a more important role within the scientific network and eventually allows them to influence the local properties of the network as they build their own team.

In 2015, Petersen published a work that offered an interesting insight into the role of ties in the formation of careers and in their evolution [116]. In his longitudinal study of careers through an egocentric perspective of the collaboration network, the author found an exponential distribution in collaboration strength, allowing to define *super ties* as ties beyond a certain extreme threshold. Such ties appear to be equally distributed across disciplines(4% of the collaborators are super ties), making long lasting partnerships an intrinsic feature of scientific collaboration. Most importantly however, super ties were shown to have a positive effect on individual careers as contributions to super ties are positively correlated with an increase in productivity in terms of numbers of publications, thus supporting the growth of careers. Similarly, publications authored by super tie collaborators are statistically more likely to attract citations on the long term, receiving on average 17% more citations, probably due to an increase in visibility brought by the presence of a super tie collaborator.

3.2.2 Centrality

From the previous subsection we have seen that as junior researchers' careers unfold into established academic positions and their early connections are carried along, they play a central role in the evolution of scientific network. But how can this property be measured? Once again, network theory comes to the rescue with the concept of *network centrality*, thanks to the computation implementation [117, 118] of basic ideas and algorithms originally introduced decades earlier in the early years of quantitative sociological studies of social networks [119, 120]. The most common type of centrality is betweenness centrality [119], which quantifies the centrality of node j by calculating the number of shortest paths between any two other nodes that goes through node j. A similar definition is the one of eigenvector centrality, which is based on a recursive idea that that a node is central in the network if it is connected to other central nodes [121]. Let a_{ij} be the adjacency matrix of a graph. The eigenvector centrality x_i of node i is given by:

$$x_i = \frac{1}{\lambda}\sum_k a_{k,i}\, x_k$$

where $\lambda \neq 0$ is a constant and $a_{i,j}$ are the elements of the adjacency matrix and λ is a constant. This score therefore recursively increases the score of a node if it is connected to other nodes with high score, with the score being eventually measured in terms of degree. This recursive equation can be solved by writing it in matrix notation and solving the eigenvector equation [122]

$$\lambda x = xA.$$

Eigenvector centrality can come in many forms [120] and is also the main idea behind Google's PageRank algorithm [123]. Regardless of the practical definition of centrality, most of the measures are found to be strongly correlated with each other, with strong values linked to a higher possibility to influence the flow of information through the network [124].

Data shows that the values of centrality in co-authorship networks are extremely skewed, with scientists with the highest score being well separated from the 2nd tier, which in turn is well separated from the 3rd and so on, thus confirming the hierarchical structure of science [125]. Also, the weighted network analysis shows that within one's collaborators, there is a strong difference in how they contribute to the short paths, with 90% of these paths going through the top 2 collaborators, therefore reinforcing the idea of strong ties between the most relevant scientists.

Centrality measures therefore represent an excellent indicator of the absolute importance of a scientist in the web of scientists, to the point where centrality itself can be shown to act as an attractor in models of preferential attachment [126]. Authors who lie in the center of network are therefore not only crucial for information spreading within the network, but also act as dominating actors who gather more attention than others to the point where the central positions allows also to have a positive effect on citations count, which are strongly correlated with centrality measures [127, 128].

3.3 Publication-based networks

In Section 2.1 we discussed the distribution of citations, which in the paper based network framework represents the analysis of the in-degree distribution. However, the structure of the connection between scientific papers can offer much more than a simple analysis of its properties. In Publication III, we focused the analysis of the connections with papers from the point of view of the community that builds around a single paper. This kind of network is called an Ego Network (EN) and it has been extensively studied in social contexts [129, 130]. In a social network where nodes are individuals, those who are part of th EN are the ones that influence the most the Ego, as they form the community in which the Ego lives. Similarly, the EN of a scientific paper is made by the set of all papers citing the Ego and of all the mutual citations between them. Fig.3.4 shows an example of an EN and of its evolution in temporal snapshots based on different time windows. The figure shows a typical pattern of the EN. The EN is initially extremely dense, with initial citers being likely to be connected to each other. The density of the EN peaks after a few years, with the building of a strongly connected core while, however, islands of isolated papers start to appear and eventually, after 5-10 years, the EN becomes extremely sparse. Interestingly, the global EN continues to grow, indicating that later papers are also citing papers from earlier windows. This indicates that, despite the original idea of the Ego being still highly considered in the scientific community, it fails to act as an aggregator of it, suggesting a specialization of the topic or, but not mutually exclusively, an increasing popularity of the ego in different disciplines.

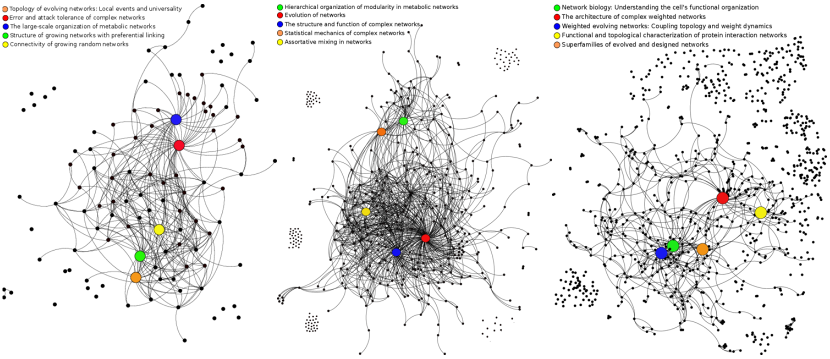
Topology of evolving networks: Local events and universality
Error and attack tolerance of complex networks
The large-scale organization of metabolic networks
Structure of growing networks with preferential linking
Connectivity of growing random networks
Hierarchical organization of modularity in metabolic networks
Evolution of networks
The structure and function of complex networks
Statistical mechanics of complex networks
Assortative mixing in networks
Network biology: Understanding the cell's functional organization
The architecture of complex weighted networks
Weighted evolving networks: Coupling topology and weight dynamics
Functional and topological characterization of protein interaction networks
Superfamilies of evolved and designed networks

Figure 3.4. Ego-network for Barabási and R. Albert's paper on scale-free networks [11]. We consider windows of size $w = 2$ at t=1 (left), t=3 (center) and t=5 (right), where t is the number of years from publication. Therefore the windows are non-overlapping and cover the intervals 1-2, 3-4 and 5-6 (years after publication). The EN is initially well connected, its link density is highest at t=3, but it quickly becomes sparse, with a growing number of isolated nodes. Some well known papers are highlighted with colors, their titles are reported at the top. Figure adapted from Publication II.

While the EN approach aims at analyzing the local structure of the community around an idea/publication and its evolution in time, it is possible to continue the analysis by "zooming out" gradually from the EN network, encompassing more and more layers of citations. Even though a single paper might not have a massive first layer (i.e. citation count), it can accumulate a vast offspring in following layers, thus spreading its influence to a large portion of the scientific network.

The growth of the influence of an idea can be studied in its evolution, assigning a stronger weight to nodes that lie in the lower circles and thus allowing to quantify the size and shape of the *wake* of a paper [131]. Interestingly, high values of this metric are able to reveal groundbreaking results that do not have high citation counts, with in particular Nobel laureates appearing as authors of some of the most significant papers. In Publication IV we found a similar pattern: we introduced a measure of the impact that a single paper has on the whole future corpus of science by allowing citing papers to "inherit" the scientific importance of the cited paper. By recursively applying the method we are thus able to measure the global contribution of a paper in the scientific network and to compare the performance of papers between citations and impact. Fig. 3.5 shows this comparison through a parameter $\delta = \frac{R_c - R_i}{R_i}$, where R_c and R_i are the rankings based on either citations (the former) or influence (the latter). δ measures the outpferformance in impact vs. citation rankings, which is extremely high for Nobel papers if compared to papers with similar citation counts, thus confirming that the cumulative importance "down the road" of scientific discoveries is not necessarily correlated to the first approximation, i.e. the citation count.

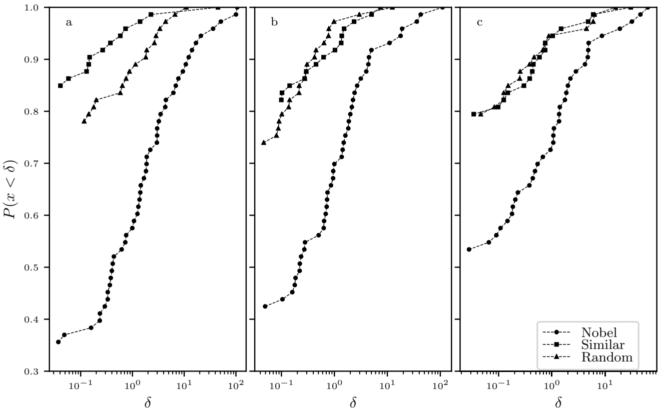

Figure 3.5. Cumulative distribution of δ for Nobel papers, paper within a 3% in citation volume in the same time interval compared to Nobel papers and for random papers after five years (panel c), ten years (panel b) and at the end of the process in 2008 (panel a). Only papers with positive δs are included. Nobel prize winning papers are more likely to climb the influence rankings, while similar papers behave similarly to random papers. Also, while the fraction of Nobel papers that is climbing the ranking is increasing as time progresses, the control group shows no significant change. Figure adapted from Publication IV.

As the previous examples show, the network structure of science can be an excellent indicator of the spread of ideas within the network. This kind of analysis has already been applied with success at a country and institutional level [132]. In this kind of framework, publications can be seen as new ideas introduced in a existing network, that are initially "exposed" to contagion from previous and become later the very source of contagion for future works. This kind of approach borrowed from epidemiology [133] is well known to be a driving force of the spread of new ideas [134] and of the emergence and diffusion of topics across disciplines. Susceptible-infected epidemic models applied to article networks show that the diffusion of new ideas over disciplines takes a long time with the incubation period ranging from 4.0 to 15.5 years [135].

Another way to look at this process is by comparison with genetics, seeing scientific ideas as genes that replicate/propagate themselves to new publications in order to survive, an idea originally introduced by Dawkins in his book *The Selfish Gene* [136]. The term he coined for these replicating entities is *meme* and it has become extremely relevant nowadays, with the explosion of similar phenomena online that behave in such a way [137]. However, as genes and viruses replicate themselves to survive, they inevitably end up competing for the same resources, thus leading to the inevitable disappearance of some of them [138]. A meme based approach to the spreading of scientific ideas has been attempted with success [139], introducing a meme score that quantifies the tendency of a scientific idea (e.g. chemical formulas or technical terms) to be replicated in a publication through a citation. Not surprisingly, high meme scores are found to be important concepts in science.

3.4 Communities, fields and multidisciplinarity

In the previous sections we talked about the global structural properties of scientific networks that can be determined from network theory. However, the opposite process can also be done. In the section on modularity and communities we discussed how the knowledge of the underlying structure of a network can be useful in order to devise methods to analyze it, similarly in science we are aware *a priori* that science is structurally organized in fields. Even within a single institution, there are separate faculties or departments, in which scientists work separate one from another, with each group focusing on different branches of science. Fields are a concept everyone is familiar with as the classical division of science in major branches such as Physics, Mathematics, Biology, Economics etc. is commonly used also outside the academic world and also the ISI has a list of 21 static fields (or rather categories) used to label all journals.

This categorization is simplistic and efficient on a superficial scale, but we know science to be a intrinsically dynamic world. Bibliometric studies [140] and studies on the co-occurrence network of scientific terms [141] have shown that fields themselves are not static, but rather follow a life-cycle that may contain branching or merging events. It appears therefore evident from these observations that also fields need to be studied not statically, but rather dynamically and that the information we know from scientific fields can be used recursively to analyze their changes in time.

Once again, works from epidemiology have been successfully applied to the topic. In a SEIR epidemic model scientists start off being Susceptible to a new idea (i.e. working in a related field), transition to being Exposed to it (i.e. they have found out about it), proceed to become Infected spreading the idea before ultimately Retiring. Empirical evidence shows that the population growth of fields can be modeled with success by this model [142].

However, these processes are not always smooth: in 1970 the philosopher T. Kuhn discussed this matter in his famous book *The Structure of Scientific Revolutions* [143], in which he described the process by which scientific knowledge progresses as being composed of periods of staticity separated by abrupt changes caused by *paradigm shifts* that challenge the scientific consensus. These shifts are mainly driven by discoveries of new information that contradicts and falsifies previous theories and methods, thus requiring collaborative effort from the scientific community in order to provide new theoretical explanations. One of the most classic examples can be seen in the foundational crisis of most scientific fields at the end of the 19th century when Darwin's evolutionary theory, Gödel's works on coherence and completeness and the new theory of Quanta caused dramatic earthquakes in Biology, Mathematics and Physics. All these events happened sharply with either the experimental observation of new phenomena or the publication of new innovative work which ultimately leads to completely new fields being born in a relative short time.

One can therefore look at structural changes in the organization of fields themselves in order to identify what are the crucial moments in the development of a single field. Studies on the temporal evolution of fields show that successful fields grow in size, becoming more dense. In particular, the relationship between the number of edges and the number of nodes follows a scaling law : edges = A(nodes)$^{\alpha}$, where A and α are constant. This process is accompanied by a topological transformation in the structure of the author network of the field: initially the authors are clustered in separate communities that, due to the densification of the network, end up merging and forming of a *large connected component* of authors, a phenomenon that does not take place for pathological cases (e.g. cold fusion in Physics) due to the innovative failure of the original idea [144]. This results show that the forming of a field is structurally connected to the forming of a sort of social network of authors around an innovative concept. This social network, shown to be dense, can therefore be used as a *ground truth* in community detection algorithms in order to identify these communities in the global network.

In fact, the changes in the connections between scientists and the subsequent change in modularity within the network can be used to accurately model the birth of new fields as a process of merging and splitting of author communities [145]. On the other hand, the diverse nature of fields and their change in time undermines the possibility to use static definition of fields as a baseline for community detection. The application of modularity maximization algorithm to paper network in fact has found that communities found in this way show a wide range of structure, varying from being strongly clustered to being barely noticeable [146]. Furthermore, fields themselves are not monolithic blocks, but rather can be organized in structured hierarchical layers; Physics for example, manifests in its own paper network a number of subfields that have different local structure, with smaller subfields being more self-referential and thus more modular [147]. This is to be expected: the larger the extent of a field (or subfield), the more it is bound to see a diversification of its ideas and the reciprocal contamination with other fields and subfields. This process leads to the birth of *interdisciplinarity* and *multidisciplinarity*.

The hierarchical nature of fields and the structural overlapping across subfields and fields has led to the necessity to use also alternative methods for community detection, such as clique percolation techniques [148]. Interdisciplinarity is not only an inevitable phenomenon of overlapping between fields, but in recent years it has shown to become an intrinsic part of the core of Physics, gradually becoming more and more relevant [149, 147]. Multidisciplainarity is slowly increasing and it can be analyzed in terms of the flow of information across fields [150], a technique that has led to the possibility of determining the stabilization of interdisciplinary fields, thus becoming new stand alone disciplines [151].

In Publication IV we studied the diffusion of scientific credit through the paper network, by spreading the scientific value of seed nodes from a field/subfield/journal of a certain year through the network. By collecting the diffused scientific value and merging it into the same groups as the seed it is possible to measure the flow of information across fields. We found that fields retain their information exponentially in time and that the exponent regulating the decay is increasing in time, thus manifesting an increase in multidisciplinarity which, however, might be a consequence of the increased rate of publication. A renormalization of time similar to the one in Publication I shows that the trend of increased interdisciplinarity is actually reversed, as shown in Fig.3.6. Interestingly, multidisciplinarity shows to be the field slowing down the most in its tendency to share information, probably as a consequence of it growing to the level of a stand-alone discipline with increased levels of self-referentiality.

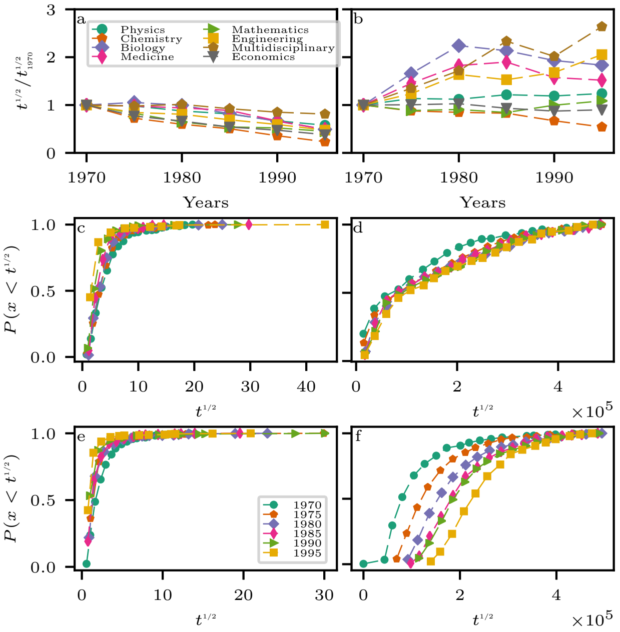

Figure 3.6. Changes in half life in time for the regular (left column, panels a-c-e) and renormalized scenario (panels b-d-f) and for different grouping of papers. Panel a shows the evolution of the half life for a number of selected fields relatively to the 1970 value, in order to compare the trend across disciplines. We can see that fields in general show a downward trend in which the half lives are decreasing. In panel b instead we can see the same evolution but but for the renormalized scenario, in which time is measured in numbers of publications published. We can see that the trend either stabilizes or is reversed. Panels c and d shows the cumulative distribution of half lives for subfields and journals for different years, while panels e and f show the same distributions with renormalized half lives. We can see that the coloring order between the two columns is reversed, indicating that also for subfields and journals are on average the same pattern as for the fields applies. Figure adapted from Publication IV.

4. Science and Metrics

In 1955, Dr. Eugene Garfield published a fundamental paper in the history of bibliometric studies [2]. In his work, Garfield introduced the idea of a citation index, i.e., a database that would allow scientists to navigate the corpus of scientific publication through citation in order to find valuable bibliographic material for their own research, an idea that eventually led to the foundation in 1960 of the Institute for Scientific Information (ISI). While advocating for the importance of such index, Garfield used as an example the possibility to quantify the number of citations: *" Thus, in the case of a highly significant article, the citation index has a quantitative value, for it may help the historian to measure the influence of the article—that is, its 'impact factor'"*, symbolically giving birth to the field of *Scientometrics*, which aims at providing a quantitative analysis of science and scientific research in general through statistical and mathematical analysis. In 1972, Garfield continued on this path by introducing a quantitative measure to rank journals based on their publication and citation count [152].

In its earliest stages the field had a huge overlap with bibliometric and library studies in general, as well as with a quantitative analysis at a micro level, such as the individual habits of scientists [153]. With the increase of the availability of data scientometrics started to differentiate as its own field aimed at the development of scientific indicators [154], also pushed by the increase need of instruments in the process of academic policy making [155], with citation based measures being the dominating base in order to assess quality in scientific output. As more citation based analysis were being introduced [156, 157], scientists also started to question the validity of such methods to assess quality of research both from a technical point of view (i.e. the mathematical validity of the methods) as well as from a philosophical one (do citations reflect quality?) [158, 159, 160, 161].

In fact, the clash between the scientific requirement to cite relevant works along with the knowledge that metrics are used in order to assess the quality of scientific research however, can lead to a vicious circle in which the methods used to analyze the scientific outputs end up influencing the selection process of cited works [162] or, in general, influencing the structure of Academia itself [163], thus compromising the previous underlying assumptions of citations as a free and voluntary choice. In spite of these limitations, citation based metrics continued being introduced and citation based rankings were introduced for authors [164] as well as for universities [165]. In this chapter I will briefly go through some of the most popular ranking measures for individual papers and authors.

4.1 Publication rankings

Even though a large of number of rankings for authors and journals were being developed, paper rankings required more time to be introduced. Unlike metrics meant for groups of papers that allow to address the rankings statistically, ranking of papers comes down to the ranking of individual nodes in a network. This task can be extremely challenging in the scientific network, especially considering the difference in citation patterns across fields both quantitatively [13] and conceptually [166]. Therefore citation counts remained for a long time a valid ranking method locally, provided that one would know what the typical citation count of a paper on a topic could be.

In order to allow for a fair ranking across *all* scientific publications instead, one would have to put into context the local properties of a paper, i.e. the community from which the citations come, with the global properties of the network, i.e. how the single community relates to all the others. This problem is closely related to what the well known Page Rank (PR) algorithm of Google does [167]. Page Rank was the most successful method among a number of solutions introduced in the 90s [168] for solving the problem of rating Web Pages in the WWW. Curiously, in their paper, Page and Brin analyze comparison between ranking pages and publications, concluding that citation counts are a far too limited tool in the presence of a large evolving network.

The idea behind PR is to provide a metric for quality of web pages that takes into account the quality of the citations themselves. In this framework therefore, a large degree (the equivalent of citation count) cannot be enough to receive a high PR as these citations might be incoming from poorly ranked nodes. In this framework therefore quality is built among a reinforcing behavior in which high quality pages "support" each other ranking wise through mutual citations or, in general, by being highly connected within the same community. Mathematically, the PR algorithm can be implemented in many ways, among which a recursive method that initially assigns equal ranking to all papers and then proceeds to propagate the ranking through the equation:

$$PR(j) = \frac{1-d}{N} + d \sum_{j \in N_i} \frac{PR(j)}{|N_j|} \tag{4.1}$$

where N is the total number of nodes and N_i is the neighborhood of node i. The PR can also be thus calculated by solving the eigenvalue equation $\vec{R} = (1-d)/N\vec{1} + dA\vec{R}$ where $\vec{R}$ is the array ranking and A is the adjacency matrix of the WWW. The possibility to express the PR algorithm in the solving of an eigenvalue equations shows that the PR is ultimately a centrality measure. The problem can be solved efficiently with the power method, requiring 52 iterations to obtain convergence for the snapshot of the WWW that Page and Bring used in 1999 [167]. The parameter d is a quantity called *damping factor* and it plays a crucial factor in the algorithm. The damping factor is linked to the implementation of the model as a random walker that propagates the PR of a single node by randomly jumping to a nearby one through its links. In this context, the damping factor represents the probability for the walker to "get bored", as the authors say, and jump to a random node in the network after $1/d$ steps on average. Practically, this factor prevents the influence of "sinks" (node or group of nodes without outgoing links) that would absorb all the rankings; with $d = 1$ we would have an infinite series of clicks, thus allowing the walker to be trapped in such sinks, while $d = 0$ would be equal to a situation in which the PR are uniform and constant. However, the damping factor also plays a fundamental role in the correct renormalization of scores across communities of different sizes [169]. If a community is strongly isolated from the core of the network (i.e. it has few incoming links), it might be difficult for the random walker to enter the community and to correctly evaluate its global PR, without the necessity to perform separate rankings.

This feature of the Page Rank thus allows to solve issues linked to different topological structures of scientific communities in citation networks both across fields and within fields [170]. In 2007 two papers attempted two adapt the Page Rank algorithm to scientific publications. Chen et al. applied the pure Page Rank algorithm to all publications belonging to the Physical Review family of journals from 1893 to 2003, with a choice of $d = 0.5$ as they believed it would better reflect the citation practices in science. Even though the PR was shown to be positively correlated with the citation count, as expected [171], a few paper were shown to be significant outliers and were identified as being important "gems" in Physics. In the same years a follow up paper came that introduced the CiteRank algorithm [172]: a generalization of the PR algorithm, in which the effects of aging into the Page Rank algorithm are taken into account. This was necessary as the PR has in intrinsic directionality based on the fact that papers cannot be cited by older ones, thus forcing the "flow" of the PR towards older entries. In the CR framework, the random walker starts from a *recent* paper and recursively follows scientific papers selecting a link not randomly, but rather in a weighed process that penalizes older papers and therefore gives a stronger value to novelty.

In Publication IV we introduced a measure that we called *persistent influence*. Despite appearing at first glance similar to PR methods, it is conceptually very different. In our approach in fact, we reversed the flow of time and we turned a stochastic process into a deterministic one. While the PR methods measure how likely a random walker is to land on a single node, we imagined a scenario in which the knowledge created in an article percolates through the network of articles. In this framework citing papers do not pass their own credit to the cited papers, but rather inherits it *from* them. Mathematically, we start from an original seed s with an initial influence $I_s = 1$ and we allow newer papers to inherit the influence through the equation :

$$I_j = \sum_{i \in N_j} \frac{I_i}{k_j^{in}} \tag{4.2}$$

where k_j^{in} is the in-degree (or, number of references) of the article j, and N_j is the set of out-neighbors. The normalization guarantees that the total influence that the cited articles have on article j is constant and that the influence value does not exceed 1. As the process continues, the influence values dilute through the network, but at the same time they are spread to increasing number of articles. At the end of the process we can then proceed to observe the influence that a single paper has had on the whole scientific network as shown in Fig.4.1.

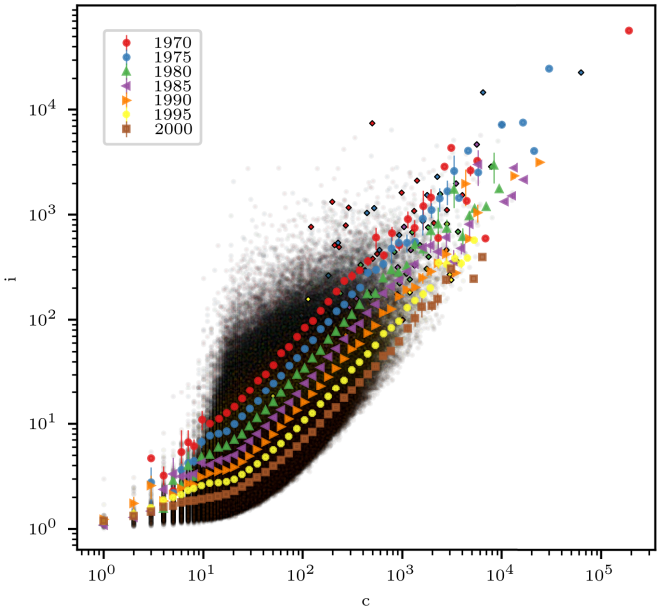

Figure 4.1. Scatter plot of values for citations vs persistent influence for different years. The full dots represent the average influence for publications within the same citation bin. Diamond shaped dots represent individual Nobel prize winning papers, the coloring of which is assigned according to the closest year to the publication date. The values appear to be correlated by a power law curve, but within each citation bin influence values can span multiple orders of magnitude. Also, Nobel prize winning papers are clustered in the top right corner, indicating both a high citation count and high influence values. Figure adapted from publication IV.

4.2 Author rankings

Science has primarily been a public endeavor carried out in public universities. As more investments were being put into research, it is no surprise that soon pressure to properly quantify scientific output would start to increase [173]. Citation counts served this purposes and have been used to decide how to allocate funds [174] as well as to select candidates for academic positions [175]. In this search for a "perfect" measure, one of the most important contributions was developed in 2005 by J.E. Hirsch [164], who introduced for the first time a clear metric aimed at ranking scientists through their citation count. The h-index is based on a very straightforward definition: an author has index h if h of their publications have gathered at least h citations each and the remaining papers have citations $\leq h$.

The new metric became immediately popular among scientists and started being considered as a standard to which to compare standard bibliometric indicators [176], both thanks to its simplicity and its ability to "rescue from obscurity" scientists who had been heavily contributing in very specific fields [177]. However, the h index was also soon discussed from a methodological point of view as authors claimed that it was not a correct way to quantify a career. In particular, it was pointed out that one can artificially alter one's index through self-citations [178] and that citations need to be weighed, as not all of them carry the same weight [179]. As other critiques followed, tackling the limitation of the h index in guaranteeing a fair ranking of scientists, new methods appeared, trying to fix the structural limitations of the h-index: indexes focusing on high cited papers (g-index) [180], indexes focusing on the average citations of the papers that grant the h-index to an author (A and AR index) [181], indexes focusing on the different volume of publications across authors (h-normalized index) [182], indexes that take into account the difference in lengths in careers (m-quotient) [183], indexes that focus only on the most cited papers (Google Scholar's *i10* index) [184] and many others [185].

As citation based indexes continued to proliferate however, another key aspect became important to tackle: what is the predictive power of the h index? Since these measures were being actively used as proxies of scientific excellence in the hiring process, it is normal to investigate the ability of the h index to predict the quality of individual careers. Hirsch himself soon tackled the aspect, reporting that the h index is able to predict a carreer: *"That is, a researcher with a high h index after 12 years is highly likely to have a high h index after* 24 *years"*[186]. While more works have similar results by combining the h index with other citation based metrics [187], other publications reported a different scenario in which past citations are only good at predicting future citations to *past* publications, but are ultimately not good at predicting future citations to *future* publications [188]. This contrast between prediction of previous results vs prediction of past results brought back the attention to the validity of the h index as a measure to predict the evolution of a career. In fact it has been argued that the h index suffers from methodological flaws due to the nature of its definition: the h index is a non stationary measure [189] which has a high auto correlation to its whole previous history, ultimately causing the h index being a good predictor of *itself* [190]. Quantitatively, any cumulative, non decreasing measure has auto correlation between its index at two different stages of the career following the relation $Cor(h(t), h(t+\Delta t)) = \sqrt{\frac{t}{t+\Delta t}}$, which means that the predictive power of such indexes is much lower when trying to estimate an individual's h-index many more years into their future than the current career academic age ($t/(t+\Delta t) \rightarrow 0$) and that for the same prediction interval (Δt) the prediction will be much more sound for a senior researcher rather than for a junior one [191]. This latter result leads to the consequence that the h index of a researcher, as their career progresses, increases regardless of their productivity [190].

These findings are ultimately in contrast with the very idea that metrics should be used to hire someone for that they will do, since such kind of citation metrics based on previous results appear to be able grasp mainly only what a scientist has done and show their strongest predictive limitations for the cases in which these will be used in real academic hiring decisions [191]. Furthermore, it can even introduce a self-reassurance bias as bureaucrats may actually take advantage of the metric auto correlation in order to have a guarantee that metrics will increase [190].

In parallel to citation based rankings however, other authors have attempted to introduce rankings based on methods similar to the Page Rank algorithm discussed in the previous paragraph [192, 193] as well as on centrality measures similar to the one mentioned in section 3.2.2 [194], but ultimately the intrinsic feasibility of the distinction between quality and quantity in scientific output is still an open question [195] and the predictability of individual indexes remains a statistical method that can possibly lead to average results, while careers have been shown to be extremely uncertain and volatile, with single events leading both to sudden career boosts [36] and negative shocks to equally extreme, yet opposite consequences [109]. Even though it is probably impossible to either develop a perfectly universal and unbiased metrics or to prevent the usage of metrics in the academic selection process, it has been argued that it would be most beneficial to minimize the increasing “taste for publication” [196] that has been gradually replacing the "taste for science" and to rely on multiple factors and measure instead of reducing the process to the evaluation of a single statistic [197].

5. Scientific Results and Discussion

5.1 Temporal patterns

Publication I studies the changes over time of the age statistics in the awarding of Nobel Prizes. In the early days of the award, prizes in Physics, Medicine and Chemistry had a $\approx 50\%$ chance to be awarded to discoveries from the previous decade, while only a smaller fraction $\approx 20\%$ of prizes was awarded to discoveries older than 20 years. In time the pattern has dramatically reversed, with nowadays more than half of the prizes being awarded over 20 years from discovery. As a result, also the age at which the Nobel prize laureates are awarded has seen a drastic increasing trend, that ultimately might lead, by the end of the century, to not be able to reward an old discovery, since the prizes cannot be awarded posthumously. While it is not simple to offer an exhaustive explanation for this trend, we suggested that a plausible one might be one of two extreme scenarios: on one hand it could be possible that the number of groundbreaking discoveries has been decreasing, therefore forcing the Nobel committee to look at older ones to find a worthy winner; on the other hand, it could be that the rate of new significant discoveries has increased so much that the limit to only 2 independent discoveries being awarded every year cannot keep up with the pace of scientific innovation.

Publication II studies the intrinsic temporal features of the life cycle of an individual paper. Publications from a dataset of over 50 million papers and 600 million citations were grouped by peak year, i.e. the year in which the higher number of yearly citations was reached, thus separating the history of a paper between its rise to "fame" and its consequent decay. In order to compare individual cycles, citation cycles were renormalized so that the maximum value (i.e. the peak) would equal to one. The time required to peak has been constantly shrinking in time across the fields of Physics, Medicine, Biology and Chemistry, with Biology showing the lowest numbers in general. The result is coherent with previous studies that show the average reference age being increasing in time, thus allowing to allocate less attention to more recent papers, which inevitably peak earlier. On the other side of the peak, the decay was found to have a form very close to either an exponential or a power law, with the former working better for older publications and the latter being a better fit as time goes by. We explained this feature as a consequence of the citation mechanism being linked to a ultradiffusive process, i.e. a mechanism in which a later event might be caused by or correlated to an earlier event or a combination of earlier events: in this case the citation count. This ultradiffusive approach allows to quantify the probability of a paper having a certain number of citations as an auto correlation function between citation counts, which can be shown analytically to be either exponential or power law in its form, as it was found in the data. Finally a non-parametric quantification of the time required to decay (i.e. an half life) allows us to show a similar pattern as for the time to peak: across fields there is a clear shrinking in the time required for a paper to be forgotten.

Publication III studies the temporal evolution of the Ego Network of highly cited scientific papers. An Ego Network built based on a single paper (the EGO) and is formed by the publications citing as nodes (the Ego is not included) with all the citations between such publications as edges. Since results of Publications I have shown that the cycle of a paper is extremely short, the EN was analyzed in its evolution in snapshots of 2 and 3 years in size, thus focusing on a temporally coherent bulk of papers that shared the Ego in their reference lists. The structure of the EN in its earliest years initially consolidates in a dense community, but is later followed by a consistent scenario, in which the networks fragment into many small components within 10 years from publication of the ego-paper, possibly linked to a specialization of the offspring of the Ego or to an increased popularity of the ego across disciplines, thus affecting the probability of cross citing.

5.2 Cumulative patterns

Publication IV studies the cumulative process of knowledge spreading stemming from the knowledge created by an individual papers. Starting from individual papers a measure called persistent influence is introduced and is based on citing papers inheriting the knowledge of cited papers. The process is then repeated recursively, thus propagating the initial influence into a cascade that eventually allows to quantify the overall influence a single paper has had on the whole corpus of scientific publications, unlike citation counts, which are based only on a local snapshot of the network limited to the first "round" of citations. Nobel winning papers are used as a benchmark for highly influential papers and in the persistent influence framework are found to be performing significantly better in their influence measures if compared to papers with similar citation counts, thus reinforcing the idea that a difference exists between local and global influence of a paper.

Publication IV also introduced a diffusive method that is used to quantify the flow of knowledge across categories (field,subfields and journals). Curves representing the loss of knowledge to other scientific categories shows a constant pattern where knowledge rapidly falls and then converges to a plateau in a typical time (the half life). While the plateau value varies across disciplines but is constant in time, the half life is decreasing in time for virtually all fields, suggesting an increase in interdisciplinarity. Furthermore, there seems to be in time a narrowing of the difference in half lives of humanistic fields (higher values) and of hard sciences (lower values), possibly linked to a structural change in the citing patterns of humanities. Multidisciplinary studies are found to have a peculiar pattern: their plateau value is increasing and their half life slowing down is among the slowest, suggesting that multidisciplinarity is possibly becoming a stand alone field that is growing internally.

Publications II and IV offer a tool of renormalization that uses cumulative information to rescale temporal patterns, thus connecting the two aspects. In both studies, temporal patterns were calculated using years as an absolute measure of time. However, in both cases, the quantities being measured were part of a system in which "updates" happen every time a new publication appears. In a system where publications come in at a constant rate, the two measures would coincide but that is not the case in science, where publications are growing at a slow, yet exponential rate. A renormalization of the time based on the number of publications instead, offers a dramatic change in the patterns observed. The speeding up in the half life for the decay of attention of a paper shown in Publication I slows down to the point where the process seems to be stable over decades and across fields, thus providing evidence for the fact that a faster decay is just a consequence of the impossibility for scientists to keep track for the ever growing amount of published material. Similarly, the speeding up of the spread of knowledge across fields found in Publication IV also changes its structure, indicating that the increasing speed of knowledge sharing across scientific fields could be explained by the increase in the speed at which the system is updated.

5.3 Discussion

Science of science as a field has seen a massive series of changes in the time since its formulation in the post war period. For a long time the pursuit of new findings in the field was hindered by the absence of properly indexed data sets that would allow a systematic analysis of the data available. As scientific data piled up over the decades and with the ever growing role of digitalization in modern times, such hinders were removed, uncovering a massive amount of information on the underlying dynamics that govern the way science works and operates.

Ever since an increasing amount of effort has been put into the uncovering of the patterns hidden in data from scientific publications: connections between papers, authors, institutions, fields, countries allowed to unravel the intrinsic properties that are at the basis of the production of scientific material. In this kind of research the basic approach has often been the one to analyze the data in locally and temporally confined snapshots. Furthermore, as scientific research sees its economical aspects become more relevant year after year, quantification of scientific output has also seen a spark in interest both from scientists and from those hiring them. This has led to a constant search for perfect metrics able to grasp universal properties for individual authors,journals or papers, compacting longitudinal careers, both past and future, into a mere number.

The research presented in this Thesis presents a diametrically opposed point of view to the matter; science does not represent a static platform for the output of new information, but is rather an ever changing system with sociological, economical and geographical characteristics, which is bound to be influenced by the constant modification of the real world on which it is ultimately based. Such changes in turn, lead to a modification of science's very own structure, thus creating patterns that are constantly evolving in time. In particular, science has been going through a constant exponential growth over the decades since the post war era, with more and more scientific knowledge accumulating on top of previous findings over a short interval of time.

The main focus of this Thesis has been to analyze these temporal and cumulative patterns both by considering their individual contribution to the analysis of scientific data as well as their united one. Only with this *combined* approach has it been possible to properly quantify the dynamics of life cycles of citation histories and Ego Network structures of individual papers, as well as the information flow between areas of science. Similarly, it allowed to introduce a paper-based measure to quantify the influence of a single publication over the whole corpus of scientific data, also allowing to track its evolution in time.

References

[1] J. W. Tukey, “Keeping research in contact with the literature: Citation indices and beyond.,” *Journal of Chemical Documentation*, vol. 2, no. 1, pp. 34–37, 1962.

[2] E. Garfield, “Citation indexes for science: A new dimension in documentation through association of ideas,” *Science*, vol. 122, no. 3159, pp. 108–111, 1955.

[3] R. E. Burton and R. W. Kebler, “The “half-life” of some scientific and technical literatures,” *American Documentation*, vol. 11, no. 1, pp. 18–22, 1960.

[4] D. de Solla Price, “Networks of scientific papers,” *Science*, vol. 149, no. 3683, pp. 510–515, 1965.

[5] P. O. Larsen and M. von Ins, “The rate of growth in scientific publication and the decline in coverage provided by science citation index,” *Scientometrics*, vol. 84, pp. 575–603, mar 2010.

[6] A. Klamer and H. P. v. Dalen, “Attention and the art of scientific publishing,” *Journal of Economic Methodology*, vol. 9, pp. 289–315, Jan. 2002.

[7] Laherrère, J. and Sornette, D., “Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales,” *Eur. Phys. J. B*, vol. 2, no. 4, pp. 525–539, 1998.

[8] Redner, S., “How popular is your paper? an empirical study of the citation distribution,” *Eur. Phys. J. B*, vol. 4, no. 2, pp. 131–134, 1998.

[9] P. S. Florence *The Economic Journal*, vol. 60, no. 240, pp. 808–810, 1950.

[10] D. de Solla Price, “A general theory of bibliometric and other cumulative advantage processes,” *Journal of the American Society for Information Science*, vol. 27, pp. 292–306, sep 1976.

[11] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *Science*, vol. 286, no. 5439, pp. 509–512, 1999.

[12] S. Redner, “Citation statistics from 110 years ofPhysical review,” *Physics Today*, vol. 58, pp. 49–54, June 2005.

[13] F. Radicchi, S. Fortunato, and C. Castellano, “Universality of citation distributions: Toward an objective measure of scientific impact,” *Proceedings of the National Academy of Sciences*, vol. 105, no. 45, pp. 17268–17272, 2008.

[14] J. King, “A review of bibliometric and other science indicators and their role in research evaluation,” *Journal of Information Science*, vol. 13, no. 5, pp. 261–276, 1987.

[15] C. Hurt, “Conceptual citation differences in science, technology, and social sciences literature,” *Information Processing & Management*, vol. 23, no. 1, pp. 1 – 6, 1987.

[16] F. Radicchi and C. Castellano, “A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions,” *PLOS ONE*, vol. 7, pp. 1–9, 03 2012.

[17] K. B. Hajra and P. Sen, “Aging in citation networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 346, no. 1–2, pp. 44 – 48, 2005.

[18] K. B. Hajra and P. Sen, “Modelling aging characteristics in citation networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 368, no. 2, pp. 575 – 582, 2006.

[19] Q. Ke, E. Ferrara, F. Radicchi, and A. Flammini, “Defining and identifying sleeping beauties in science,” *Proceedings of the National Academy of Sciences*, vol. 112, no. 24, pp. 7426–7431, 2015.

[20] P. W. Higgs, “Broken symmetries and the masses of gauge bosons,” *Phys. Rev. Lett.*, vol. 13, pp. 508–509, Oct 1964.

[21] A. COLLABORATION, “Observation of a new particle in the search for the standard model higgs boson with the {ATLAS} detector at the {LHC},” *Physics Letters B*, vol. 716, no. 1, pp. 1 – 29, 2012.

[22] J. Ellis, M. K. Gaillard, and D. V. Nanopoulos, “A historical profile of the higgs boson,” 2012.

[23] J. Ellis, M. K. Gaillard, and D. Nanopoulos, “A phenomenological profile of the higgs boson,” *Nuclear Physics B*, vol. 106, pp. 292 – 340, 1976.

[24] S. L. De Groote and J. L. Dorsch, “Measuring use patterns of online journals and databases,” *J Med Libr Assoc*, vol. 91, pp. 231–240, Apr 2003.

[25] M. J. Stringer, M. Sales-Pardo, and L. A. Nunes Amaral, “Effectiveness of journal ranking schemes as a tool for locating information,” *PLOS ONE*, vol. 3, pp. 1–8, 02 2008.

[26] J. A. Evans, “Electronic publication and the narrowing of science and scholarship,” *Science*, vol. 321, no. 5887, pp. 395–399, 2008.

[27] A. Verstak, A. Acharya, H. Suzuki, S. Henderson, M. Iakhiaev, C. C. Lin, and N. Shetty, “On the shoulders of giants: The growing impact of older articles,” *CoRR*, vol. abs/1411.0275, 2014.

[28] R. K. Pan, A. M. Petersen, F. Pammolli, and S. Fortunato, “The memory of science: Inflation, myopia, and the knowledge network,” *CoRR*, vol. abs/1607.05606, 2016.

[29] C. Tenopir, D. W. King, S. Edwards, and L. Wu, “Electronic journals and changes in scholarly article seeking and reading patterns,” *Aslib Proceedings*, vol. 61, no. 1, pp. 5–32, 2009.

[30] H. D. White, B. Wellman, and N. Nazer, “Does citation reflect social structure?: Longitudinal evidence from the “globenet” interdisciplinary research group,” *Journal of the American Society for Information Science and Technology*, vol. 55, no. 2, pp. 111–126, 2004.

[31] O. Persson, W. Glänzel, and R. Danell, “Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies,” *Scientometrics*, vol. 60, no. 3, pp. 421–432, 2004.

[32] W. Glänzel and B. Thijs, “Does co-authorship inflate the share of self-citations?,” *Scientometrics*, vol. 61, no. 3, pp. 395–404, 2004.

[33] G. Wolfgang, T. Bart, and S. Balázs, “A bibliometric approach to the role of author self-citations in scientific communication,” *Scientometrics*, vol. 59, no. 1, pp. 63–77, 2004.

[34] J. H. Fowler and D. W. Aksnes, “Does self-citation pay?,” *Scientometrics*, vol. 72, no. 3, pp. 427–437, 2007.

[35] M. L. Wallace, V. Larivière, and Y. Gingras, “A small world of citations? the influence of collaboration networks on citation practices,” *PLOS ONE*, vol. 7, pp. 1–10, 03 2012.

[36] A. Mazloumian, Y.-H. Eom, D. Helbing, S. Lozano, and S. Fortunato, “How citation boosts promote scientific paradigm shifts and nobel prizes,” *PLOS ONE*, vol. 6, pp. 1–6, 05 2011.

[37] C. J. Lee, C. R. Sugimoto, G. Zhang, and B. Cronin, “Bias in peer review,” *Journal of the American Society for Information Science and Technology*, vol. 64, no. 1, pp. 2–17, 2013.

[38] R. K. Merton, “Thein Science,” *Science*, vol. 159, pp. 56–63, Jan. 1968.

[39] L. Bornmann and H. Daniel, “What do citation counts measure? a review of studies on citing behavior,” *Journal of Documentation*, vol. 64, no. 1, pp. 45–80, 2008.

[40] A.-L. Barabasi and Z. N. Oltvai, “Network biology: understanding the cell’s functional organization,” *Nat Rev Genet*, vol. 5, pp. 101–113, Feb 2004.

[41] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabasi, “The large-scale organization of metabolic networks,” *Nature*, vol. 407, pp. 651–654, Oct 2000.

[42] L. A. N. Amaral, A. Scala, M. Barthélémy, and H. E. Stanley, “Classes of small-world networks,” *Proc Natl Acad Sci U S A*, vol. 97, pp. 11149–11152, Oct 2000. 200327197[PII].

[43] H. Zhu, X. Wang, and J.-Y. Zhu, “Effect of aging on network structure,” *Phys. Rev. E*, vol. 68, p. 056121, Nov 2003.

[44] R. Ghosh and B. A. Huberman, “Information relaxation is ultradiffusive,” *arXiv:1310.2619 [physics]*, Oct. 2013. arXiv: 1310.2619.

[45] M. Wang, G. Yu, and D. Yu, “Measuring the preferential attachment mechanism in citation networks,” *Physica A: Statistical Mechanics and its Applications*, vol. 387, no. 18, pp. 4692 – 4698, 2008.

[46] Q. L. Burrel, “Stochastic modelling of the first-citation distribution,” *Scientometrics*, vol. 52, no. 1, pp. 3–12, 2001.

[47] M. L. Wallace, V. Larivière, and Y. Gingras, “Modeling a century of citation distributions,” *Journal of Informetrics*, vol. 3, no. 4, pp. 296 – 303, 2009.

[48] M. E. J. Newman, “The first-mover advantage in scientific publication,” *EPL (Europhysics Letters)*, vol. 86, no. 6, p. 68001, 2009.

[49] Y.-H. Eom and S. Fortunato, “Characterizing and modeling citation dynamics,” *PLOS ONE*, vol. 6, pp. 1–7, 09 2011.

[50] K.-I. Goh and A.-L. Barabási, “Burstiness and memory in complex systems,” *EPL (Europhysics Letters)*, vol. 81, no. 4, p. 48002, 2008.

[51] M. Karsai, M. Kivelä, R. K. Pan, K. Kaski, J. Kertész, A.-L. Barabási, and J. Saramäki, “Small but slow world: How network topology and burstiness slow down spreading,” *Phys. Rev. E*, vol. 83, p. 025102, Feb 2011.

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

[53] C. Goffman, “And what is your erdos number?,” *The American Mathematical Monthly*, vol. 76, no. 7, pp. 791–791, 1969.

[54] P. Cayley, “On the analytical forms called trees,” *American Journal of Mathematics*, vol. 4, no. 1/4, p. 266, 1881.

[55] D. König, *Theory of Finite and Infinite Graphs*. Birkhäuser, 1990.

[56] J.-C. Fournier, *Théorie des graphes et applications avec exercices et problèmes revue et augmentée*. Hermes Science Publications.

[57] *Graph Theory and Theoretical Physics*. Academic Press Inc, 1968.

[58] R. D. Luce and A. D. Perry, “A method of matrix analysis of group structure,” *Psychometrika*, vol. 14, no. 2, pp. 95–116, 1949.

[59] R. S. Weiss and E. Jacobson, “A method for the analysis of the structure of complex organizations,” *American Sociological Review*, vol. 20, no. 6, pp. 661–668, 1955.

[60] P. Erdös and A. Rényi, “On random graphs, I,” *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959.

[61] J. E. Cohen, “Threshold phenomena in random structures,” *Discrete Applied Mathematics*, vol. 19, no. 1, pp. 113 – 128, 1988.

[62] M. Altmann, “Susceptible-infected-removed epidemic models with dynamic partnerships,” *J Math Biol*, vol. 33, no. 6, pp. 661–675, 1995.

[63] M. J. Keeling, “The effects of local spatial structure on epidemiological invasions,” *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 266, no. 1421, pp. 859–867, 1999.

[64] B. K. Fosdick, D. B. Larremore, J. Nishimura, and J. Ugander, “Configuring random graph models with fixed degree sequences,” 2016. arXiv: 1608.00607.

[65] E. F. Connor and D. Simberloff, “The assembly of species communities: Chance or competition?,” *Ecology*, vol. 60, p. 1132, dec 1979.

[66] M. Gail and N. Mantel, “Counting the number of r × c contingency tables with fixed margins,” *Journal of the American Statistical Association*, vol. 72, p. 859, dec 1977.

[67] D. J. Watts and S. H. Strogatz, “Collective dynamics of /‘small-world/’ networks,” *Nature*, vol. 393, pp. 440–442, Jun 1998.

[68] R. Albert and A.-L. Barabási, “Statistical mechanics of complex networks,” *Reviews of Modern Physics*, vol. 74, pp. 47–97, jan 2002.

[69] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási, “Structure and tie strengths in mobile communication networks,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7332–7336, 2007.

[70] T. A. Davis and Y. Hu, “The university of florida sparse matrix collection,” *ACM Trans. Math. Softw.*, vol. 38, pp. 1:1–1:25, Dec. 2011.

[71] P. Holme and J. Saramäki, “Temporal networks,” *Physics Reports*, vol. 519, no. 3, pp. 97 – 125, 2012. Temporal Networks.

[72] M. Kivela, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, “Multilayer networks,” *Journal of Complex Networks*, vol. 2, pp. 203–271, jul 2014.

[73] S. Boccaletti, G. Bianconi, R. Criado, C. del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin, "The structure and dynamics of multilayer networks," *Physics Reports*, vol. 544, pp. 1–122, nov 2014.

[74] M. Szell, R. Lambiotte, and S. Thurner, "Multirelational organization of large-scale social networks in an online world," *Proceedings of the National Academy of Sciences*, vol. 107, no. 31, pp. 13636–13641, 2010.

[75] E. N. Gilbert, "Random graphs," *Ann. Math. Statist.*, vol. 30, pp. 1141–1144, 12 1959.

[76] A.-L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nat Rev Genet*, vol. 5, pp. 101–113, Feb 2004.

[77] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, pp. 661–703, nov 2009.

[78] M. E. J. Newman, "Assortative mixing in networks," *Phys. Rev. Lett.*, vol. 89, p. 208701, Oct 2002.

[79] M. E. J. Newman, "Mixing patterns in networks," *Physical Review E*, vol. 67, feb 2003.

[80] D. S. Callaway, J. E. Hopcroft, J. M. Kleinberg, M. E. J. Newman, and S. H. Strogatz, "Are randomly grown graphs really random?," *Physical Review E*, vol. 64, sep 2001.

[81] R. Noldus and P. Van Mieghem, "Assortativity in complex networks," *Journal of Complex Networks*, vol. 3, no. 4, p. 507, 2015.

[82] J. P. Sterbenz, D. Hutchison, E. K. Çetinkaya, A. Jabbar, J. P. Rohrer, M. Schöller, and P. Smith, "Resilience and survivability in communication networks: Strategies, principles, and survey of disciplines," *Computer Networks*, vol. 54, no. 8, pp. 1245 – 1265, 2010. Resilient and Survivable networks.

[83] G. Fagiolo, "Clustering in complex directed networks," *Physical Review E*, vol. 76, aug 2007.

[84] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, jan 2003.

[85] S. Milgram, "The small-world problem," *Psychology Today*, vol. 1, no. 1, 1967.

[86] W. W. Zachary, "An information flow model for conflict and fission in small groups," *Journal of Anthropological Research*, vol. 33, no. 4, pp. 452–473, 1977.

[87] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3–5, pp. 75 – 174, 2010.

[88] E. Ravasz, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, pp. 1551–1555, aug 2002.

[89] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[90] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," 2006.

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

[92] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *Proceedings of the National Academy of Sciences*, vol. 104, no. 1, pp. 36–41, 2007.

[93] A. Lancichinetti, M. Kivelä, J. Saramäki, and S. Fortunato, "Characterizing the community structure of complex networks," *PLOS ONE*, vol. 5, pp. 1–8, 08 2010.

[94] A. Lancichinetti and S. Fortunato, "Limits of modularity maximization in community detection," *Phys. Rev. E*, vol. 84, p. 066122, Dec 2011.

[95] B. H. Good, Y.-A. de Montjoye, and A. Clauset, "Performance of modularity maximization in practical contexts," *Phys. Rev. E*, vol. 81, p. 046106, April 2010.

[96] S. Fortunato and D. Hric, "Community detection in networks: A user guide," *Physics Reports*, vol. 659, pp. 1 – 44, 2016. Community detection in networks: A user guide.

[97] L. Peel, D. B. Larremore, and A. Clauset, "The ground truth about metadata and community detection in networks," *Science Advances*, vol. 3, p. e1602548, may 2017.

[98] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences*, vol. 98, no. 2, pp. 404–409, 2001.

[99] A. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A: Statistical Mechanics and its Applications*, vol. 311, no. 3–4, pp. 590 – 614, 2002.

[100] C. S. Wagner and L. Leydesdorff, "Network structure, self-organization, and the growth of international collaboration in science," *Research Policy*, vol. 34, no. 10, pp. 1608 – 1618, 2005.

[101] D. de Solla Price and S. Gürsey, "Studies in scientometrics i transience and continuance in scientific authorship," *Ciência da Informação*, vol. 4, no. 1, 1975.

[102] M. E. Newman, *Who Is the Best Connected Scientist?A Study of Scientific Coauthorship Networks*, pp. 337–370. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004.

[103] S. Uddin, L. Hossain, and K. Rasmussen, "Network effects on scientific collaborations," *PLOS ONE*, vol. 8, no. 2, pp. 1–12, 2013.

[104] G. Palla, A.-L. Barabasi, and T. Vicsek, "Quantifying social group evolution," *Nature*, vol. 446, pp. 664–667, April 2007.

[105] P. Deville, D. Wang, R. Sinatra, C. Song, V. D. Blondel, and A.-L. Barabási, "Career on the move: Geography, stratification, and scientific impact," *Scientific Reports*, vol. 4, pp. 4770 EP –, Apr 2014. Article.

[106] J. Hoekman, K. Frenken, and R. J. Tijssen, "Research collaboration at a distance: Changing spatial patterns of scientific collaboration within europe," *Research Policy*, vol. 39, no. 5, pp. 662 – 673, 2010. Special Section on Government as Entrepreneur.

[107] L. Leydesdorff and C. S. Wagner, "International collaboration in science and the formation of a core group," *Journal of Informetrics*, vol. 2, no. 4, pp. 317 – 325, 2008.

[108] K. Kaplan, "Academia: The changing face of tenure," *Nature*, vol. 468, pp. 123–125, nov 2010.

[109] A. M. Petersen, M. Riccaboni, H. E. Stanley, and F. Pammolli, "Persistence and uncertainty in the academic career," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 5213–5218, mar 2012.

[110] R. Guimera, "Team assembly mechanisms determine collaboration network structure and team performance," *Science*, vol. 308, pp. 697–702, apr 2005.

[111] A. Pentland, "The new science of building great teams.," *Harv Bus Rev*, vol. 90, pp. 60–69, 2012.

[112] R. K. Pan and J. Saramäki, "The strength of strong ties in scientific collaboration networks," *EPL (Europhysics Letters)*, vol. 97, p. 18007, jan 2012.

[113] Q. Ke and Y.-Y. Ahn, "Tie strength distribution in scientific collaboration networks," *Physical Review E*, vol. 90, sep 2014.

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

[115] A. M. Petersen, W.-S. Jung, J.-S. Yang, and H. E. Stanley, "Quantitative and empirical demonstration of the matthew effect in a study of career longevity," *Proceedings of the National Academy of Sciences*, vol. 108, no. 1, pp. 18–23, 2011.

[116] A. M. Petersen, "Quantifying the impact of weak, strong, and super ties in scientific careers," *Proceedings of the National Academy of Sciences*, vol. 112, no. 34, pp. E4671–E4680, 2015.

[117] M. J. Newman, "A measure of betweenness centrality based on random walks," *Social Networks*, vol. 27, no. 1, pp. 39 – 54, 2005.

[118] U. Brandes, "A faster algorithm for betweenness centrality," *The Journal of Mathematical Sociology*, vol. 25, no. 2, pp. 163–177, 2001.

[119] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, no. 1, pp. 35–41, 1977.

[120] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[121] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *The Journal of Mathematical Sociology*, vol. 2, no. 1, pp. 113–120, 1972.

[122] B. Ruhnau, "Eigenvector-centrality — a node-centrality?," *Social Networks*, vol. 22, no. 4, pp. 357 – 365, 2000.

[123] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," tech. rep., Stanford InfoLab, 1999.

[124] T. W. Valente, K. Coronges, C. Lakon, and E. Costenbader, "How correlated are network centrality measures?," *Connect (Tor)*, vol. 28, pp. 16–26, Jan 2008. 20505784[pmid].

[125] M. E. J. Newman, "Scientific collaboration networks. II. shortest paths, weighted networks, and centrality," *Physical Review E*, vol. 64, jun 2001.

[126] A. Abbasi, L. Hossain, and L. Leydesdorff, "Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks," *Journal of Informetrics*, vol. 6, no. 3, pp. 403 – 412, 2012.

[127] A. Abbasi, J. Altmann, and L. Hossain, "Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures," *Journal of Informetrics*, vol. 5, no. 4, pp. 594 – 607, 2011.

[128] E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer, "Predicting scientific success based on coauthorship networks," *EPJ Data Science*, vol. 3, no. 1, p. 9, 2014.

[129] J. J. McAuley and J. Leskovec, "Learning to discover social circles in ego networks.," in *NIPS*, vol. 2012, pp. 548–56, 2012.

[130] V. Arnaboldi, M. Conti, A. Passarella, and F. Pezzoni, "Analysis of ego network structure in online social networks," in *Privacy, security, risk and trust (PASSAT), 2012 international conference on and 2012 international confernece on social computing (SocialCom)*, pp. 31–40, IEEE, 2012.

[131] D. F. Klosik and S. Bornholdt, "The citation wake of publications detects nobel laureates' papers," *PLOS ONE*, vol. 9, pp. 1–9, 12 2014.

[132] K. Börner, S. Penumarthy, M. Meiss, and W. Ke, "Mapping the diffusion of scholarly knowledge among major u.s. research institutions," *Scientometrics*, vol. 68, no. 3, pp. 415–426, 2006.

[133] N. A. Christakis and J. H. Fowler, "Social contagion theory: examining dynamic social networks and human behavior," *Statistics in Medicine*, vol. 32, pp. 556–577, jun 2012.

[134] L. M. Bettencourt, A. Cintrón-Arias, D. I. Kaiser, and C. Castillo-Chávez, "The power of a good idea: Quantitative modeling of the spread of ideas from epidemiological models," *Physica A: Statistical Mechanics and its Applications*, vol. 364, pp. 513 – 536, 2006.

[135] I. Z. Kiss, M. Broom, P. G. Craze, and I. Rafols, "Can epidemic models describe the diffusion of topics across disciplines?," *Journal of Informetrics*, vol. 4, no. 1, pp. 74 – 82, 2010.

[136] R. Dawkins, *The Selfish Gene.* Oxford University Press, 1976.

[137] J. Leskovec, L. Backstrom, and J. Kleinberg, "Meme-tracking and the dynamics of the news cycle," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, (New York, NY, USA), pp. 497–506, ACM, 2009.

[138] L. Weng, A. Flammini, A. Vespignani, and F. Menczer, "Competition among memes in a world with limited attention," *Scientific Reports*, vol. 2, mar 2012.

[139] T. Kuhn, M. c. v. Perc, and D. Helbing, "Inheritance patterns in citation networks reveal scientific memes," *Phys. Rev. X*, vol. 4, p. 041036, Nov 2014.

[140] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong, "TextFlow: Towards better understanding of evolving topics in text," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, pp. 2412–2421, dec 2011.

[141] D. Chavalarias and J.-P. Cointet, "Phylomemetic patterns in science evolution—the rise and fall of scientific fields," *PLOS ONE*, vol. 8, pp. 1–11, 02 2013.

[142] L. M. A. Bettencourt, D. I. Kaiser, J. Kaur, C. Castillo-Chávez, and D. E. Wojick, "Population modeling of the emergence and development of scientific fields," *Scientometrics*, vol. 75, no. 3, p. 495, 2008.

[143] T. S. Kuhn, *The structure of scientific revolutions*. Chicago: University of Chicago Press, 1970.

[144] L. M. Bettencourt, D. I. Kaiser, and J. Kaur, "Scientific discovery and topological transitions in collaboration networks," *Journal of Informetrics*, vol. 3, no. 3, pp. 210 – 221, 2009. Science of Science: Conceptualizations and Models of Science.

[145] X. Sun, J. Kaur, S. Milojević, A. Flammini, and F. Menczer, "Social dynamics of science," *Scientific Reports*, vol. 3, jan 2013.

[146] P. Chen and S. Redner, “Community structure of the physical review citation network,” *Journal of Informetrics*, vol. 4, no. 3, pp. 278 – 290, 2010.

[147] R. Sinatra, P. Deville, M. Szell, D. Wang, and A.-L. Barabási, “A century of physics,” *Nature Physics*, vol. 11, pp. 791–796, oct 2015.

[148] M. Herrera, D. C. Roberts, and N. Gulbahce, “Mapping the evolution of scientific fields,” *PLoS ONE*, vol. 5, p. e10355, may 2010.

[149] R. K. Pan, S. Sinha, K. Kaski, and J. Saramäki, “The evolution of interdisciplinarity in physics research,” *Scientific Reports*, vol. 2, aug 2012.

[150] A. L. Porter and I. Rafols, “Is science becoming more interdisciplinary? measuring and mapping six research fields over time,” *Scientometrics*, vol. 81, pp. 719–745, apr 2009.

[151] M. Rosvall and C. T. Bergstrom, “Mapping change in large networks,” *PLOS ONE*, vol. 5, pp. 1–7, 01 2010.

[152] E. Garfield, “Citation analysis as a tool in journal evaluation: Journals can be ranked by frequency and impact of citations for science policy studies,” *Science*, vol. 178, pp. 471–479, nov 1972.

[153] B. Latour and S. Woolgar, *Laboratory Life: The Construction of Scientific Facts, 2nd Edition*. Princeton University Press, 1986.

[154] L. Leydesdorff and S. Milojević, “Scientometrics,” 2012. arXiv: 1208.4566.

[155] B. R. Martin, “Foresight in science and technology,” *Technology Analysis & Strategic Management*, vol. 7, no. 2, pp. 139–168, 1995.

[156] J. R. Cole and S. Cole, “The ortega hypothesis: Citation analysis suggests that only a few scientists contribute to scientific progress,” *Science*, vol. 178, pp. 368–375, oct 1972.

[157] A. F. J. van Raan, “Advanced bibliometric methods to assess research performance and scientific development: basic principles and recent practical applications,” *Research Evaluation*, vol. 3, pp. 151–166, dec 1993.

[158] P. O. Seglen, “The skewness of science,” *Journal of the American Society for Information Science*, vol. 43, no. 9, pp. 628–638, 1992.

[159] M. H. MacRoberts and B. R. MacRoberts, “Problems of citation analysis: A critical review,” *Journal of the American Society for Information Science*, vol. 40, no. 5, pp. 342–349, 1989.

[160] M. H. MacRoberts and B. R. MacRoberts, “Problems of citation analysis: A critical review,” *Journal of the American Society for Information Science*, vol. 40, pp. 342–349, sep 1989.

[161] P. O. Seglen, “Citations and journal impact factors: questionable indicators of research quality,” *Allergy*, vol. 52, pp. 1050–1056, nov 1997.

[162] L. L. Hargens and H. Schuman, “Citation counts and social comparisons: Scientists’ use and evaluation of citation index data,” *Social Science Research*, vol. 19, no. 3, pp. 205 – 221, 1990.

[163] A. Siow, “Tenure and other unusual personnel practices in academia,” *Journal of Law, Economics, & Organization*, vol. 14, no. 1, pp. 152–173, 1998.

[164] J. E. Hirsch, “An index to quantify an individual’s scientific research output,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 46, pp. 16569–16572, 2005.

[165] S. R. Consultancy, "The academic ranking of world universities." `http://www.shanghairanking.com/`.

[166] M. Franceschet, "The difference between popularity and prestige in the sciences and in the social sciences: A bibliometric analysis," *Journal of Informetrics*, vol. 4, pp. 55–63, jan 2010.

[167] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," Technical Report 1999-66, Stanford InfoLab, November 1999.

[168] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, pp. 604–632, sep 1999.

[169] H. Xie, K.-K. Yan, and S. Maslov, "Optimal ranking in networks with community structure," *Physica A: Statistical Mechanics and its Applications*, vol. 373, pp. 831–836, jan 2007.

[170] S. Maslov and S. Redner, "Promise and pitfalls of extending google's PageRank algorithm to citation networks," *Journal of Neuroscience*, vol. 28, pp. 11103–11105, oct 2008.

[171] S. Fortunato, M. Boguñá, A. Flammini, and F. Menczer, "Algorithms and models for the web-graph," ch. Approximating PageRank from In-Degree, pp. 59–71, Berlin, Heidelberg: Springer-Verlag, 2008.

[172] D. Walker, H. Xie, K.-K. Yan, and S. Maslov, "Ranking scientific publications using a model of network traffic," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2007, pp. P06010–P06010, jun 2007.

[173] J. Lane and S. Bertuzzi, "Measuring the results of science investments," *Science*, vol. 331, pp. 678–680, feb 2011.

[174] L. Bornmann and H.-D. Daniel, "Selection of research fellowship recipients by committee peer review. reliability, fairness and predictive validity of board of trustees' decisions," *Scientometrics*, vol. 63, pp. 297–320, apr 2005.

[175] K. W. Boyack and K. Börner, "Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers," *Journal of the American Society for Information Science and Technology*, vol. 54, pp. 447–461, jan 2003.

[176] A. F. J. van Raan, "Comparison of the hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups," *Scientometrics*, vol. 67, pp. 491–502, jun 2006.

[177] P. Ball, "Index aims for fair ranking of scientists," *Nature*, vol. 436, pp. 900–900, aug 2005.

[178] A. PURVIS, "The h index: playing the numbers game," *Trends in Ecology & Evolution*, vol. 21, pp. 422–422, aug 2006.

[179] M. C. Wendl, "H-index: however ranked, citations need context," *Nature*, vol. 449, pp. 403–403, sep 2007.

[180] L. Egghe, "Theory and practise of the g-index," *Scientometrics*, vol. 69, pp. 131–152, oct 2006.

[181] B. Jin, L. Liang, R. Rousseau, and L. Egghe, "The r- and AR-indices: Complementing the h-index," *Chinese Science Bulletin*, vol. 52, pp. 855–863, mar 2007.

[182] A. Sidiropoulos, D. Katsaros, and Y. Manolopoulos, "Generalized hirsch h-index for disclosing latent facts in citation networks," *Scientometrics*, vol. 72, pp. 253–280, jun 2007.

[183] Q. L. Burrell, "On the h-index, the size of the hirsch core and jin's a-index," *Journal of Informetrics*, vol. 1, pp. 170–177, apr 2007.

[184] Google, "Google scholar citations open to all." https://scholar.googleblog.com/2011/11/google-scholar-citations-open-to-all.html.

[185] S. Alonso, F. Cabrerizo, E. Herrera-Viedma, and F. Herrera, "h-index: A review focused in its variants, computation and standardization for different scientific fields," *Journal of Informetrics*, vol. 3, pp. 273–289, oct 2009.

[186] J. E. Hirsch, "Does the h index have predictive power?," *Proceedings of the National Academy of Sciences*, vol. 104, pp. 19193–19198, nov 2007.

[187] D. E. Acuna, S. Allesina, and K. P. Kording, "Future impact: Predicting scientific success," *Nature*, vol. 489, pp. 201–202, sep 2012.

[188] A. Mazloumian, "Predicting scholars' scientific impact," *PLoS ONE*, vol. 7, p. e49246, nov 2012.

[189] O. Penner, A. M. Petersen, R. K. Pan, and S. Fortunato, "Commentary: The case for caution in predicting scientists' future impact," *Physics Today*, vol. 66, pp. 8–9, apr 2013.

[190] M. Schreiber, "How relevant is the predictive power of the h-index? a case study of the time-dependent hirsch index," *Journal of Informetrics*, vol. 7, pp. 325–329, apr 2013.

[191] O. Penner, R. K. Pan, A. M. Petersen, K. Kaski, and S. Fortunato, "On the predictability of future impact in science," *Scientific Reports*, vol. 3, oct 2013.

[192] F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani, "Diffusion of scientific credits and the ranking of scientists," *Phys. Rev. E*, vol. 80, p. 056103, Nov 2009.

[193] E. Yan, Y. Ding, and C. R. Sugimoto, "P-rank: An indicator measuring prestige in heterogeneous scholarly networks," *Journal of the American Society for Information Science and Technology*, pp. n/a–n/a, 2010.

[194] E. Yan and Y. Ding, "Applying centrality measures to impact analysis: A coauthorship network analysis," *Journal of the American Society for Information Science and Technology*, vol. 60, pp. 2107–2118, oct 2009.

[195] J. Kaur, E. Ferrara, F. Menczer, A. Flammini, and F. Radicchi, "Quality versus quantity in scientific impact," *Journal of Informetrics*, vol. 9, pp. 800–808, oct 2015.

[196] M. Osterloh, "Governance by numbers. does it really work in research?," *Analyse & Kritik*, vol. 32, jan 2010.

[197] B. S. Frey and K. Rost, "Do rankings reflect research quality?," *Journal of Applied Economics*, vol. 13, pp. 1–38, may 2010.

Aalto-DD 226/2017

9 789526 077253

ISBN 978-952-60-7725-3 (printed)
ISBN 978-952-60-7726-0 (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

DOCTORAL
DISSERTATIONS