Aalto University publication series
DOCTORAL DISSERTATIONS /

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 XX, at a public examination held at the lecture hall XX of the school on X January 2017 at XX.

Aalto University
School of Science
Department of Computer Science

Supervising professor
Kimmo Kaski

Thesis advisors
Mikko Kivelä
Santo Fortunato

Preliminary examiners
Alexander Petersen,
UC Merced
School of Engineering 2
Suite 315 5200 N. Lake Road
Merced, CA 95343

Angel Sánchez, Prof
Departamento de Matemáticas and Institute UC3M-BS for Financial Big Data
Universidad Carlos III de Madrid,
Av. de la Universidad, 30, 28911 Leganés, Madrid, Spain

Aalto University publication series
DOCTORAL DISSERTATIONS /

ISBN (printed)
ISBN (pdf)
ISSN-L 1799-4934
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:

Unigrafia Oy
Helsinki

Finland

441 697
Printed matter

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 /

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 pa[illegible]. It shows that typical patterns in citation count and diffusion of information between fie[illegible]eavily 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)	**ISBN (pdf)**	
ISSN-L 1799-4934	**ISSN (printed)** 1799-4934	**ISSN (pdf)** 1799-4942
Location of publisher Helsinki	**Location of printing** Helsinki	**Year**
Pages 100	**urn** http://urn.fi/URN:ISBN:	

Preface

Espoo, September 22, 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 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

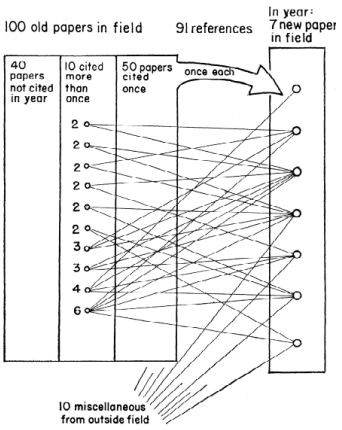

Figure 1.1. Representation of citations as a network structure. Figure adapted from [4] with permission of The American Association for the Advancement of Science.

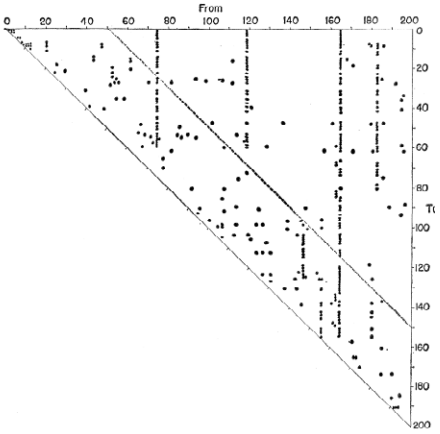

Figure 1.2. Representation of citations as an adjacency matrix. Figure adapted from [4] with permission of The American Association for the Advancement of Science.

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

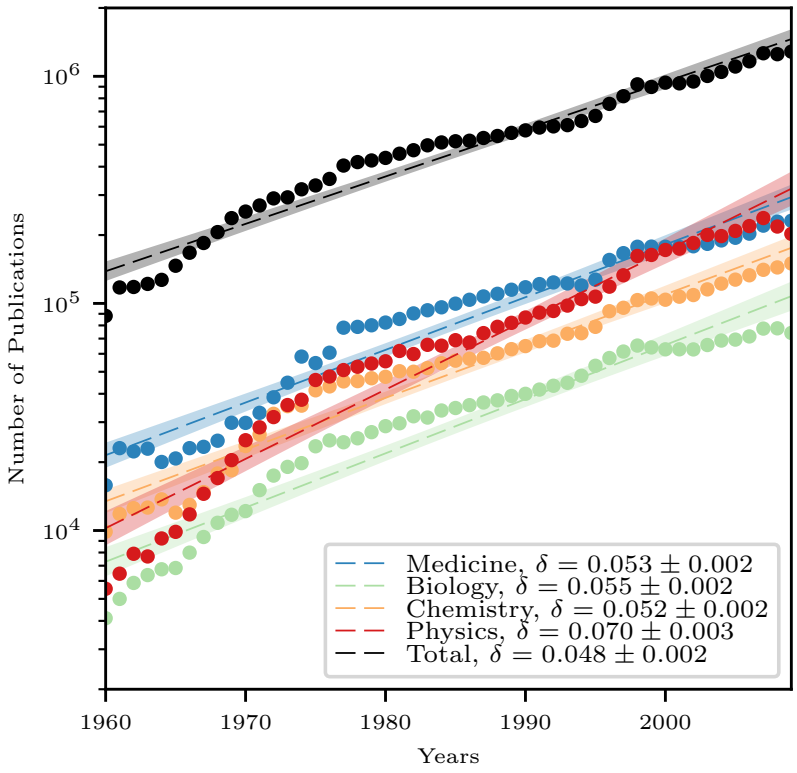

Figure 1.3. Growth of publication in science and for a selected number of fields based on our ISI dataset of over 50 million publications and 600 million citations. The rate of growth can be well approximated by an exponential curve. Figure adapted from Publication II.

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.

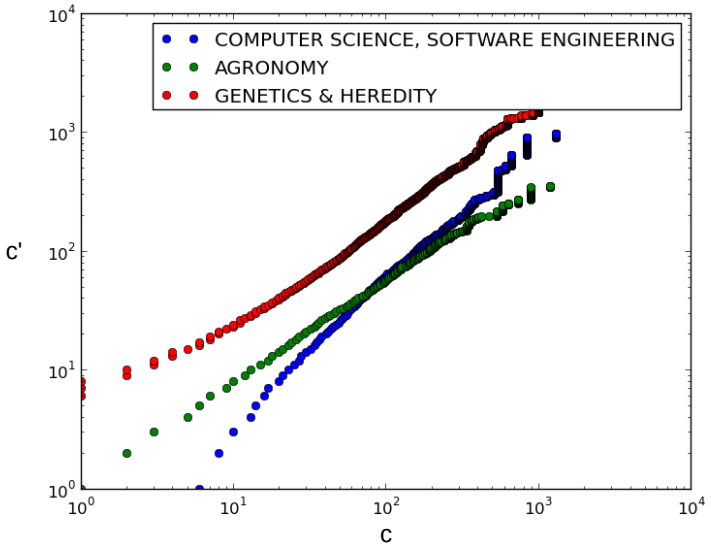

Figure 2.1. c' vs c adapted from [16] and reproduced with our data set. We can see that the scaling follows the relation: $c' = ac^{\alpha}$ where a is a pre-factor and α is a field dependent scaling factor.

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.

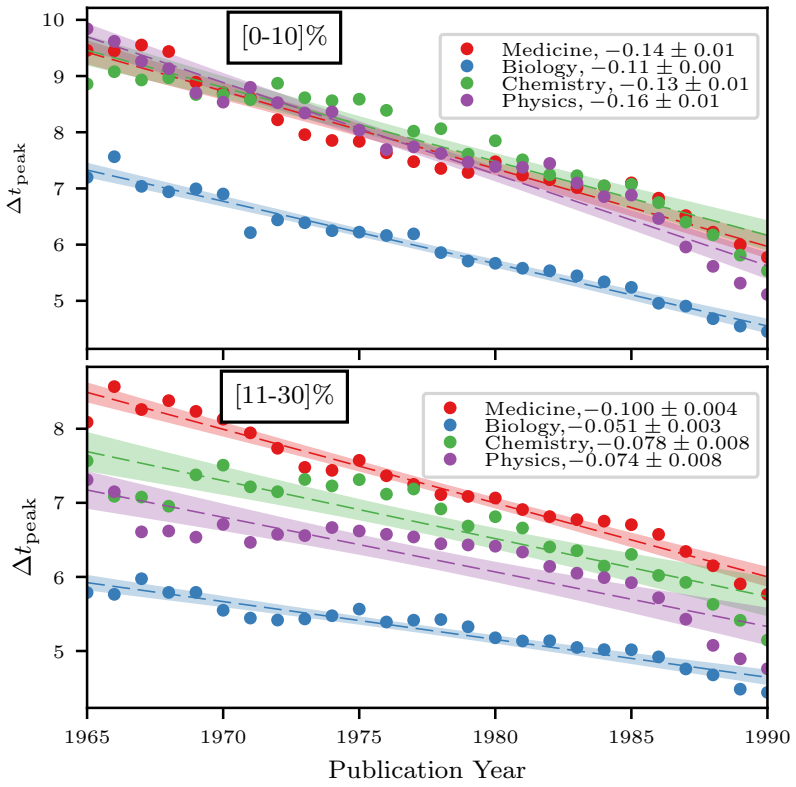

Figure 2.2. Time evolution of the mean values of time to peak Δt_{peak} for top 10% (top) and [11-30]% percentiles (bottom) of our ISI dataset. Δt_{peak} represents the time elapsed between the publication of a paper and the year in which it reached its maximum yearly citation count. The mean value $\langle \Delta t_{\text{peak}} \rangle$ 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

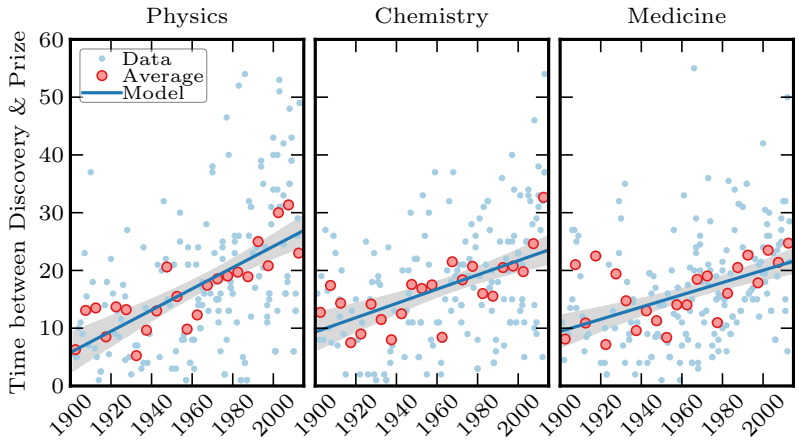

Figure 2.3. Time lag between discovery and Nobel prize vs year in which the prize was awarded for Physics, Chemistry and Medicine, created with data from Publication I. For each Nobel prize we searched bibliographic material on the author in order to identify one or more publications that could be directly 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 $\approx$ 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.

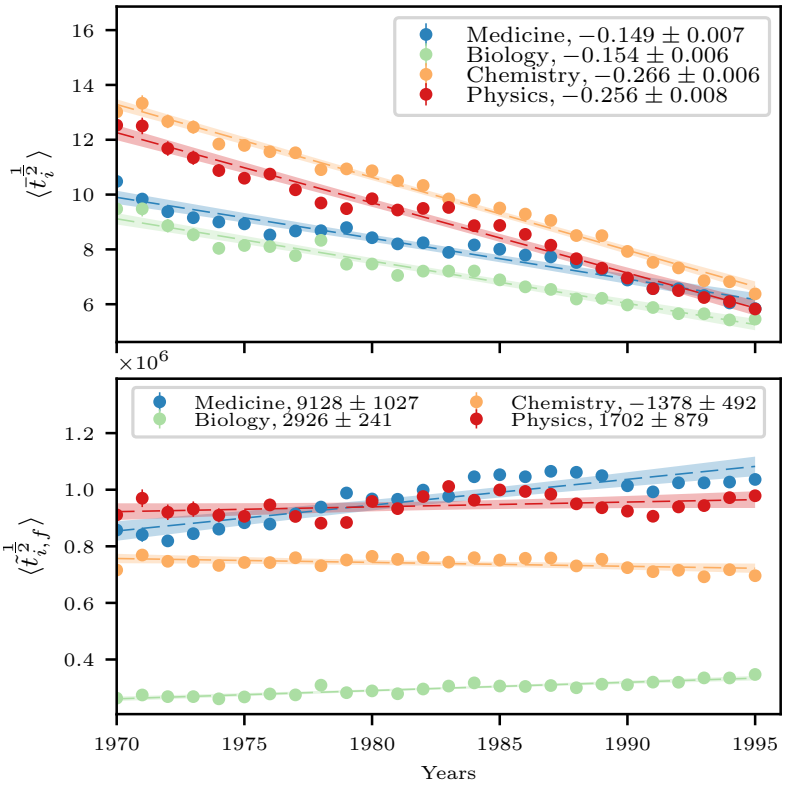

Figure 2.4. The evolution of the half life of papers after the peak $\langle t_i^{\frac{1}{2}} \rangle$ in terms of absolute time (top) and $\langle \tilde{t}_{i,f}^{\frac{1}{2}} \rangle$ in terms of the number of publications (bottom) for the four different fields and for the top 10% percentile. For 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.

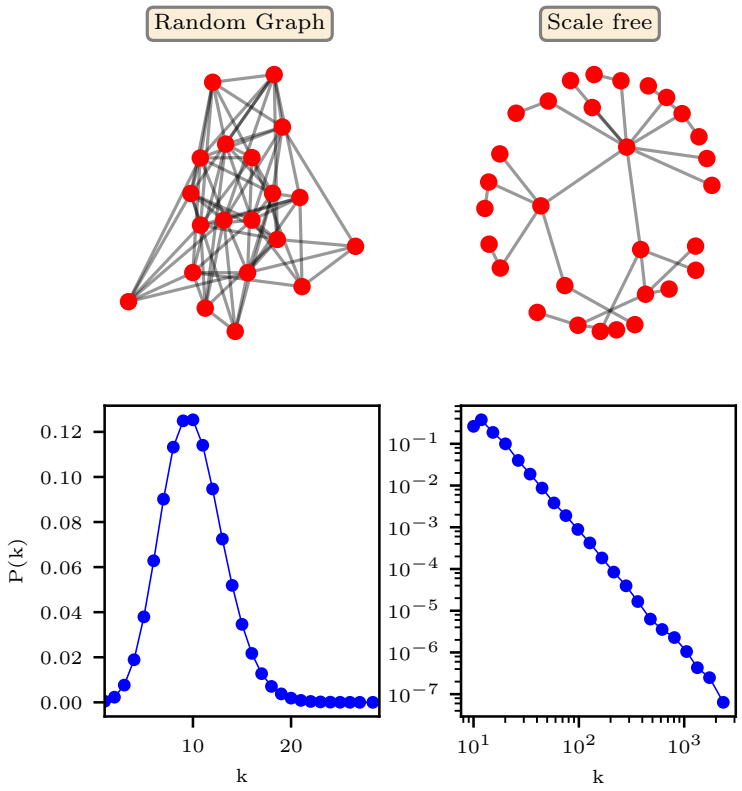

Figure 3.1. Difference in topology and degree distribution between a random graph (left) and a scale-free network (right). The random network has its degree distribution heavily centered around its average, with no significant outliers. In the scale-free model instead, degrees can span multiple orders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

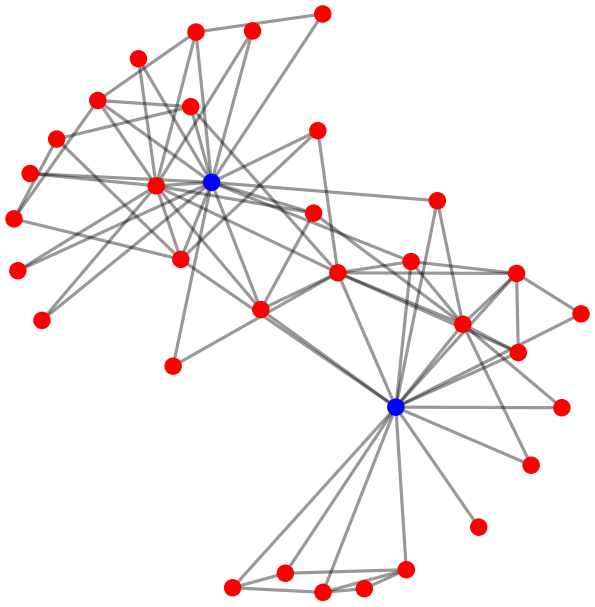

Figure 3.2. Visualization of the karate network based on the data from [86]. The network is visibly structured around the two hubs colored in blue, with clustered communities around each hub and a few nodes acting as intermediaries between the two communities.

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].

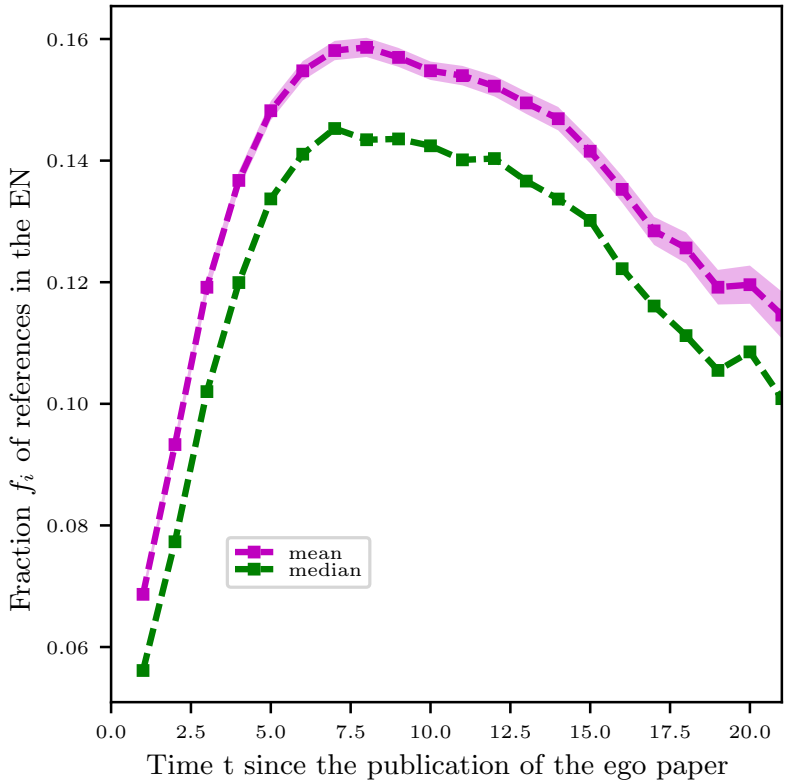

Figure 3.3. Time evolution of the mean and median of the fraction f_i of references of papers of the full Ego Network belonging to the Ego Network as a function of the number of years since publication. In this framework, f_i is the EN equivalent of the modularity of the community that is formed around 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.

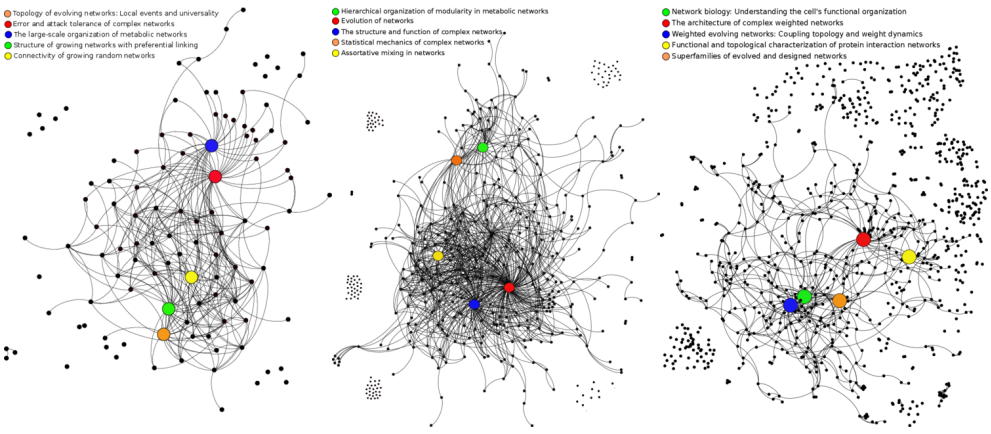

Figure 3.4. Ego-network for Barabási and R. Albert's paper on scale-free networks [11]. We consider windows of size $w = 2$ at t=1 (left), t=3 (center) and t=5 (right), where t is the number of years from publication. Therefore the windows are non-overlapping and cover the intervals 1-2, 3-4 and 5-6 (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.

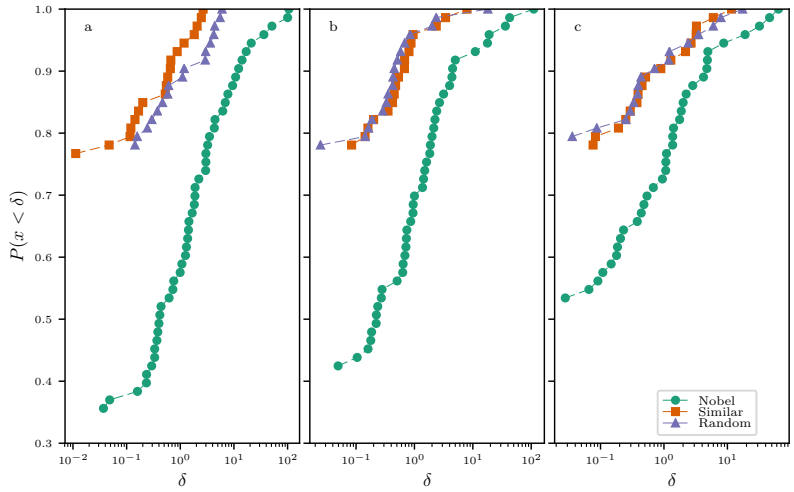

Figure 3.5. Cumulative distribution of δ for Nobel papers, paper within a 3% in citation volume in the same time interval compared to Nobel papers and for random papers after five years (panel c), ten years (panel b) and at the end of the process in 2008 (panel a). Only papers with positive δs are included. Nobel prize winning papers (green dots) are more likely to climb the influence rankings, while similar papers (orange) dots behave similarly to random papers (purple dots). 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.

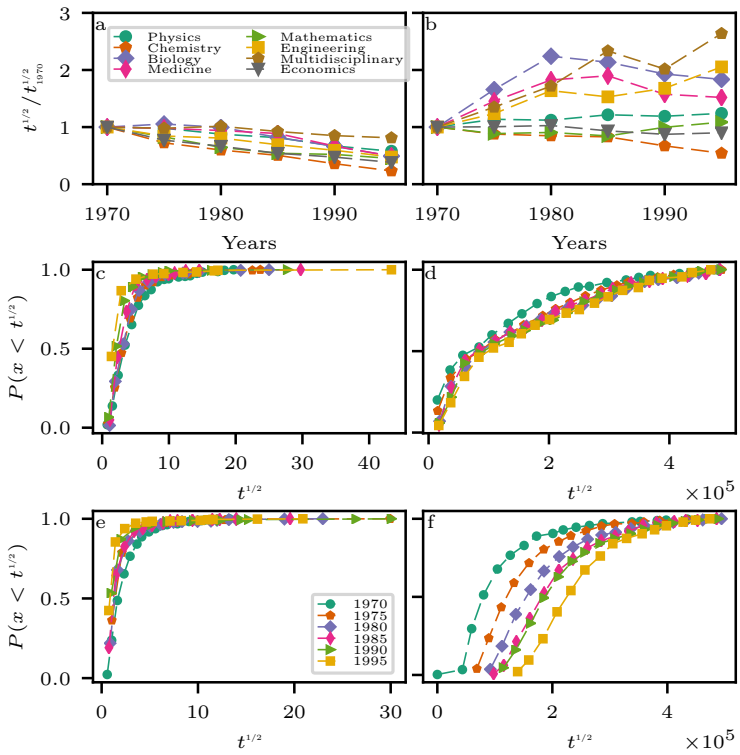

Figure 3.6. Changes in half life in time for the regular (left column, panels a-c-e) and renormalized scenario (panels b-d-f) and for different grouping of papers. Panel a shows the evolution of the half life for a number of selected fields relatively to the 1970 value, in order to compare the trend 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.

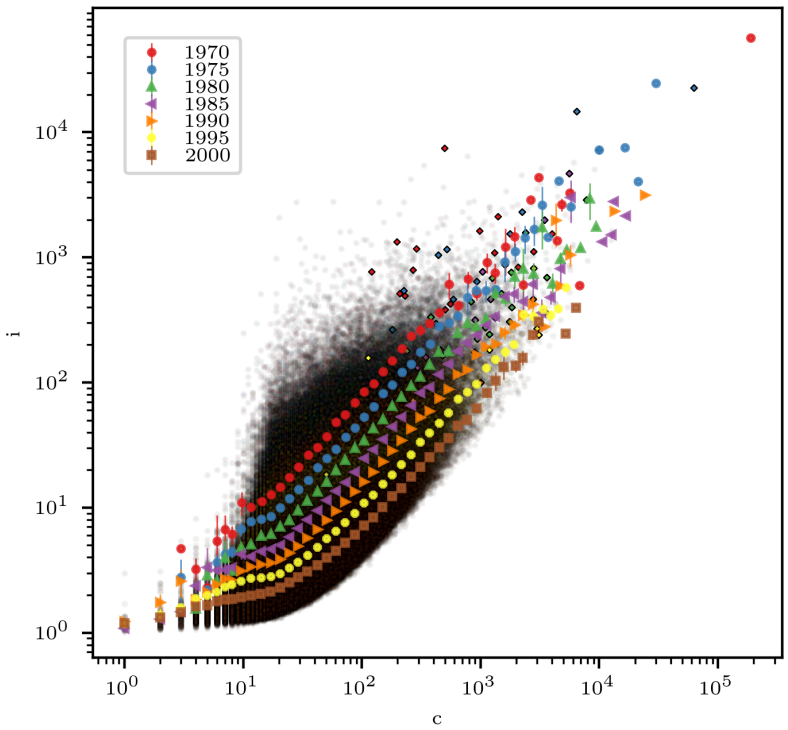

Figure 4.1. Scatter plot of values for citations vs persistent influence for different years. The full dots represent the average influence for publications within the same citation bin. Diamond shaped dots represent individual Nobel prize winning papers, the coloring of which is assigned according to 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 ($t/(t+\Delta t) \to 0$) 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.

Publication 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.

The Nobel Prize delay

Francesco Becattini,[1] Arnab Chatterjee,[2] Santo Fortunato,[2] Marija Mitrović,[2] Raj Kumar Pan,[2] and Pietro Della Briotta Parolo[2]

[1] *Universitá di Firenze and INFN Sezione di Firenze, Florence, Italy*
[2] *Department of Biomedical Engineering and Computational Science, Aalto University School of Science, P.O. Box 12200, FI-00076, Finland*

The time lag between the publication of a Nobel discovery and the conferment of the prize has been rapidly increasing for all disciplines, especially for Physics. Does this mean that science is running out of groundbreaking discoveries or that, on the contrary, there have been too many breakthroughs?

The 2013 Nobel Prize in Physics was awarded to Higgs and Englert for their prediction of the existence of the Higgs boson. Though the Higgs particle was experimentally discovered at CERN in 2012, the original theoretical works date back to the 1960s. Thus, it took about half a century of intense work to confirm their prediction.

Long time lags between discovery and recognition are not unusual. In fact, it has been significantly increasing over the years (Figure 1). Let $\Delta^{\mathrm{D}\to\mathrm{N}}$ be the time between

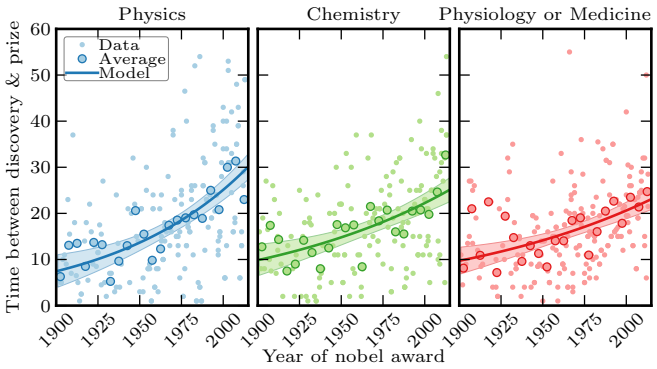

FIG. 1. Time difference (in years) between the discovery and the awarding of the Nobel prize, versus the year when the award is received. Each plot shows the raw data, the 5-year average, and the exponential fit with its confidence interval. The lag is increasing for the three fields, with rates of 0.012±0.002, 0.008±0.002 and 0.008±0.001 for Physics, Chemistry and Physiology or Medicine, respectively.

the discovery and the Nobel award. We model the variation of $\Delta^{\mathrm{D}\to\mathrm{N}}$ with time t by considering an exponential law:

$$\Delta^{\mathrm{D}\to\mathrm{N}}(t) = c_\alpha \exp(\alpha t), \tag{1}$$

where α is the rate of increase in $\Delta^{\mathrm{D}\to\mathrm{N}}$ and c_α is a proportionality constant. Figure 1 shows an increase in $\Delta^{\mathrm{D}\to\mathrm{N}}$ for all fields. The predicted values and indicated 95% confidence intervals are given by the exponential regression model. Using linear regression we get consistent results. The rate of increase in $\Delta^{\mathrm{D}\to\mathrm{N}}$ is highest for Physics, followed by Chemistry and by Physiology or Medicine. On the x-axis of Fig. 1 we report the year when the Nobel Prize is actually awarded. This means that future awards for already published discoveries will have no influence on the ones shown in our plots, they will contribute to the future evolution of the curves.

Figure 2 elaborates the details of the field-specific $\Delta^{\mathrm{D}\to\mathrm{N}}$-dynamics. It shows the percentage of prizes awarded over 20 years of the discovery. The predicted values and indicated 95% confidence intervals are given by logistic polynomial regressions. Here we estimate

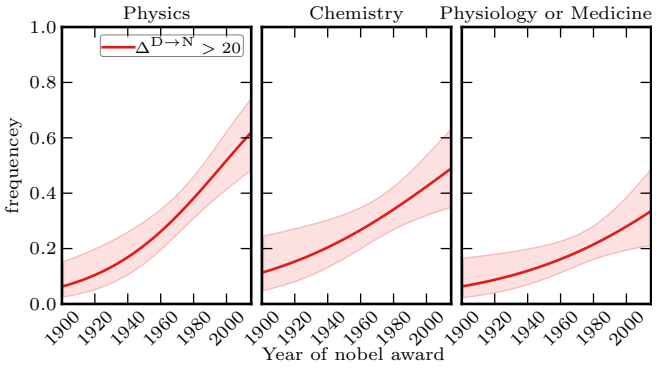

FIG. 2. The frequency of prizes awarded over 20 years of the discovery is increasing for all disciplines. The growth is fastest for Physics and slowest for Physiology or Medicine.

first-degree logistic polynomial regressions for all the fields. The conditional probability of the discovery being awarded within T year is given by

$$\Pr(\Delta^{\mathrm{D}\to\mathrm{N}} < T|t) = \frac{1}{1+\exp[-(\mu+\nu t)]}, \tag{2}$$

where the parameters μ and ν are estimated using the maximum likelihood method. After 1985, about 15% of Physics, 18% of Chemistry and 9% of Physiology or Medicine prizes are awarded within 10 years of their discovery. In contrast, before 1940 about 61% of Physics, 48% of Chemistry and 45% of Physiology or Medicine prizes are awarded within 10 years of the discovery. Correspondingly, after 1985 about 60% of Physics, 52% of Chemistry and 49% of Physiology or Medicine prizes are awarded over 20 years of the discovery. In comparison, before 1940 only about 11% of Physics, 15% of Chemistry and 24% of Physiology or Medicine prizes were awarded over 20 years of the discovery. In all fields the frequency of the prize being awarded over 20 years since discovery

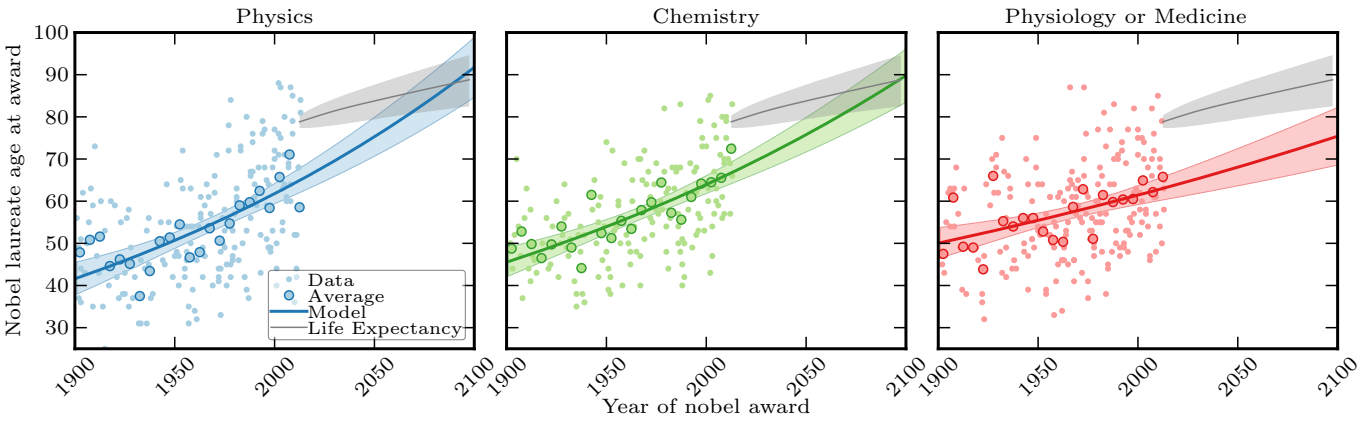

FIG. 3. Change in the age of the scientist at which the Nobel prize is awarded. For all fields there is an increasing trend. For Physics and Chemistry the rate of increase is similar (0.0040 ± 0.0005 and 0.0034 ± 0.0004), while for Physiology or Medicine the increase is much smaller (0.0020 ± 0.0005). The progression of the average life expectancy in the United States is also shown in grey.

is increasing. The rate of increase in the frequency of getting the award 20 plus years since the discovery is fastest for Physics and slowest for Physiology or Medicine.

As a result of the increasing time to recognize a Nobel discovery, the age at which laureates receive the award is also increasing. We consider how the age at which scientists are awarded the Nobel prize a^{N} is changing with time. An exponential increase is represented by

$$a^{\mathrm{N}}(t) = c_\gamma \exp(\gamma t), \tag{3}$$

where γ is the rate of increase of the age and c_γ is a proportionality constant.

Figure 3 shows that a^{N} is increasing for the three fields. We also used the regression model to project the age of the laureates at the time of the award until the end of the century. The predicted values and indicated 95% confidence intervals are given by the exponential regression model. The figure also shows the projected life expectancies (of men and women combined together) across the 21st century. Here we used the data of the United States as a proxy of the life expectancy (as US citizens hasve been awarded the majority of Nobel prizes). The expectancy is based on WPP2012 estimates using the medium scenario and the 95% prediction interval is also shown [1]. We found that by the end of this century for the fields of Physics and Chemistry, the Nobel laureates' age at discovery would become higher than the life expectancy. Therefore, if this trend is maintained, by the end of this century it might become technically impossible to confer the Nobel prize, as it is not possible to award it posthumously.

What is the reason of the increasing delay between discovery and recognition? A plausible explanation could be that the frequency of groundbreaking discoveries is decreasing. Interestingly, since no more than two discoveries can be awarded with the Nobel prize in the same year, it could even be that there are too many important discoveries, and that, in order not to lose worthy winners, one is forced to dig deeper and deeper in the past. Also, in many cases it takes much longer now than before to verify a groundbreaking result (e.g., 48 years in the case of the Higgs boson). All the above generally applies to any discipline. Yet the delay is increasing much faster for Physics than for Medicine. This seems to confirm the common feeling of an increasing time needed to achieve new discoveries in basic natural sciences, a somewhat worrisome trend.

DATA

We collected data on dates of birth, the year of Nobel prizes and year(s) of publication(s) of prize winning work. As a primary data source we used the Nobel Foundation's website, `nobelprize.org`. In the cases where the information was not sufficient to accurately identify year(s) of prize winning publication we consulted all the publications of the Nobel Laureates using `google.scholar.com`. We then determined the year of the most relevant publication related to the topic of the Nobel prize award. We also consulted the biographies of the laureates and other resources, such as `nobel.caltech.edu/`, `journals.aps.org/prl/50years/milestones`.

[1] United Nations, *World Population Prospects: The 2012 Revision* (Department of Economic and Social Affairs, Population Division, New York, 2013).

Publication 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.

Journal of Informetrics 9 (2015) 734–745

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Attention decay in science

Pietro Della Briotta Parolo[a], Raj Kumar Pan[a,*], Rumi Ghosh[b], Bernardo A. Huberman[c], Kimmo Kaski[a], Santo Fortunato[a]

[a] Complex Systems Unit, Aalto University School of Science, P.O. Box 12200, FI-00076, Finland
[b] Robert Bosch LLC, Palo Alto, CA 94304, USA
[c] Mechanisms and Design Lab, Hewlett Packard Enterprise Labs, Palo Alto, CA, USA

ARTICLE INFO

Article history:
Received 6 March 2015
Received in revised form 14 July 2015
Accepted 14 July 2015
Available online 1 September 2015

Keywords:
Decay of attention
Citation count
Time evolution

ABSTRACT

The exponential growth in the number of scientific papers makes it increasingly difficult for researchers to keep track of all the publications relevant to their work. Consequently, the attention that can be devoted to individual papers, measured by their citation counts, is bound to decay rapidly. In this work we make a thorough study of the life-cycle of papers in different disciplines. Typically, the citation rate of a paper increases up to a few years after its publication, reaches a peak and then decreases rapidly. This decay can be described by an exponential or a power law behavior, as in ultradiffusive processes, with exponential fitting better than power law for the majority of cases. The decay is also becoming faster over the years, signaling that nowadays papers are forgotten more quickly. However, when time is counted in terms of the number of published papers, the rate of decay of citations is fairly independent of the period considered. This indicates that the attention of scholars depends on the number of published items, and not on real time.

1. Introduction

Scientific publications in peer reviewed journals serve as the standard medium through which most of the progress of science is recorded. Besides offering a mechanism for claiming priorities and exposing results to be checked by others, publishing is also a way to attract attention of other scientists working on related problems. Attention, measured by the number and lifetime of citations, is the main currency of the scientific community, and along with other forms of recognition forms the basis for promotions and the reputation of scientists (Petersen et al., 2014). As Franck (Franck, 1999), Klamer and van Dalen (Klamer & Dalen, 2002) have pointed out, there is an attention economy at work in science, in which those seeking attention through the production of new knowledge are rewarded by being cited by their peers, whose own standing is measured by the amount of citations they receive.

The attention economy is also at work in many other fields besides science, ranging from entertainment to marketing, and is responsible for the phenomenon of stars, i.e., people whose income in attention far exceeds the norm in their own endeavors. Moreover, attention is a strong motivator of productivity. Recently, it has been shown that the productivity of YouTube videos exhibits a strong positive dependence on the attention they receive, measured by the number of downloads (Huberman, Romero, & Wu, 2009). Conversely, a lack of attention leads to a decrease in the number of videos uploaded and the consequent drop in productivity, which in many cases asymptotes to no uploads whatsoever.

* Corresponding author.
E-mail address: rajkumar.pan@aalto.fi (R.K. Pan).

http://dx.doi.org/10.1016/j.joi.2015.07.006
1751-1577/

Table 1
Basic statistics of the different scientific fields we considered: Clinical Medicine, Molecular Biology, Chemistry and Physics. They represent the most active fields in terms of the total volume of publications. Here, N_P is the number of publications in a given field, c_{max} is the maximum number of citations to a given paper in that field and $\langle c \rangle$ is the average number of citations to all the papers in that field.

Field	N_P	c_{max}	$\langle c \rangle$
Clinical Medicine	10833626	25604	11
Molecular Biology	2849144	296498	24
Chemistry	4565197	134441	14
Physics	5583183	31759	13

Decision making and marketing, among others, are based on the mechanisms ruling how attention is stimulated and maintained (Dukas, 2004; Kahneman, 1973; Pashler, 1998; Pieters, Rosbergen, & Wedel, 1999; Reis, 2006). Over the past years, thanks to the Internet, a huge amount of data has allowed a thorough investigation of the dynamics of collective attention to online content, ranging from news stories (Dezsö et al., 2006; Ghosh & Huberman, 2014; Wu & Huberman, 2007), to videos (Crane & Sornette, 2008) and memes (Leskovec, Backstrom, & Kleinberg, 2009; Matsubara, Sakurai, Prakash, Li, & Faloutsos, 2012; Weng, Flammini, Vespignani, & Menczer, 2012). Here attention is measured by the number of users' views, visits, posts, downloads, tweets. It is also noted that the attention decays over time, not only because novelty fades, but also because the human capacity to pay attention to new content is limited. A typical temporal pattern is characterized by an initial rapid growth, followed by a decay. The decay turns out to be slower than exponential: power law fits give the best results, stretched exponentials being preferable in particular cases (Wu & Huberman, 2007).

In this paper we focus on the decay of attention in science, on the basis of scientific articles, which like any other content, become obsolete after a while. Typically this happens because their results are surpassed by those of successive papers, which then "steal" attention from them. The problem of the obsolescence of scientific contents has received a lot of attention in scientometrics. The typical approach is to study the evolution of the number of citations received by a paper in a given time frame (usually one year), since its publication. The nature of the decay has been controversial, between claims of an exponential trend (Avramescu, 1979; Medo, Cimini, & Gualdi, 2011; Nakamoto, 1988) and analyses supporting a slower power law curve (Bouabid, 2011; Bouabid & Larivière, 2013; Pollman, 2000; Redner, 2005). This is partly due to the different types of analysis and the use of distinct data sources. Note that patterns of individual papers are usually noisy, as one cannot count on the high statistics available for online contents: the number of tweets posted on a single popular topic may exceed the total number of scientific publications ever made.

On the other hand, in contrast to online sources, bibliographic databases enable one to perform a longitudinal study of the life cycles of papers. In this work we make a systematic analysis of papers' life cycles, across different scientific fields and historical periods. We find that the decay of attention for individual papers can be described both by exponential and power law behaviors. Exponential fits turn out to be preferable in the majority of cases. These results are compatible with a relaxation of attention modeled by ultradiffusion, as observed for the popularity of online content (Ghosh & Huberman, 2014). We also found that attention is dying out more rapidly with time. However, due to the ongoing exponential growth of scientific publications, which is known to influence citation patterns (Egghe, 2000; Yang, Ma, Song, & Qiu, 2010), we conjecture that the faster decay observed nowadays is a consequence of the much larger pool of papers among which attention has to be distributed. In fact, if time is renormalized in terms of the number of papers published in the corresponding period (e.g., in each given year), we find that the rescaled curves die out at comparable rates across the decades.

2. Material and methods

2.1. Data description

Our data set consists of all publications (articles and reviews) written in English till the end of 2010 included in the database of the Thomson Reuters (TR) Web of Science. For each publication we extracted its year of publication, the subject category of the journal in which it is published and the corresponding citations to that publication. Based on the subject category of the journal (determined by TR) of the publication, the papers were categorized in broader disciplines such as Physics, Medicine, Chemistry and Biology (see Table 1). Most analyses are carried out using the top 10% papers (based on their total number of citations), as it allows to include a sufficient number of papers from older times, but still keeping the number of yearly citations large enough to allow for a statistically valid analysis. The analysis of papers with relatively lower citations follow qualitatively similar behavior and is shown in the Appendix.

2.2. Data fitting and F-statistics

We measure the trend in the temporal evolution of the different plots using the least square method. We consider the *F*-statistics for a significant linear regression relationship between the response variable and the predictor variable. We used it to compare the statistical models that best fit the population from which the data were sampled. As the *F*-score takes into account both the number of data points available for the fit and the number of degrees of freedom of the model, it is possible to compare the accuracy of the fit for different models with different parameters or between data sets of different size.

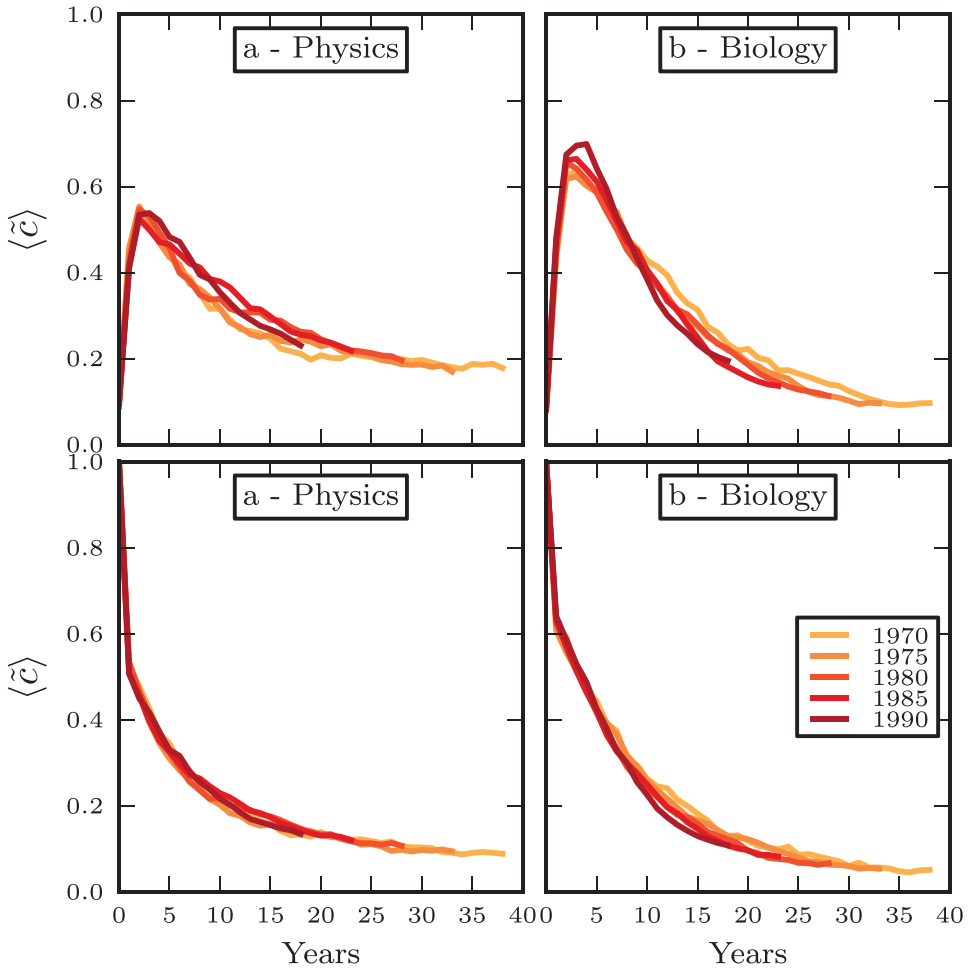

Fig. 1. The citation life-cycle is both field dependent and time dependent. (Top) Normalized number of citations per year received by papers in Physics and Biology published in the same year, for different publication years. Normalization is done by dividing the number of citations by the peak value reached by the paper. (Bottom) The decay in the (normalized) citation trajectory of papers in both fields after the peak year. For both disciplines, the averaged citation trajectories are calculated for papers in the top decile (top 10%) based on their total number of citations.

3. Results and discussions

3.1. Evolution of the number of citations

We first look at the way citations received by a paper change with time. Since different scientific fields are characterized by different volumes of publications and citations, many features of the citation trajectory are field dependent. However, for most fields the number of yearly citations $c_i(t)$ to a given paper i rises after its publication and peaks within 2–7 years. The peak is followed by a decay in the number of citations that reflects the obsolescence of older knowledge. Fig. 1 (top panels) shows the normalized citation trajectory $\tilde{c}_i(t) \equiv c_i(t)/c_i^{\max}$ of papers in Physics and Biology. Here, $c_i^{\max}$ is the maximum number of citations received by paper i in any given year after its publication. Fig. 2 shows a summary of the renormalization process and different measures used for analysis. For both disciplines, the citation trajectories of papers published over different years show systematic changes with time. New papers have higher citation rates for the first few years, whereas over longer periods of time old papers have higher citation rates. Some irregularity in the tail of the citation trajectories might be due to the heterogeneity in the time to reach the peak number of citations Δt_{peak}. The change in the citation rate over time is more evident when we group the papers based on their *peak year*, i.e., year in which they receive the maximum number of citations. Thus, the peak year represents the year in which a paper is at the peak of its attention. Fig. 1 (bottom panels) show that the decay pattern is more robust when the papers were aggregated according to their peak year as compared to their publication year. This is true for other groups of papers as well: Appendix Fig. B.1 shows the same pattern for the papers in the [11–30] percentile.

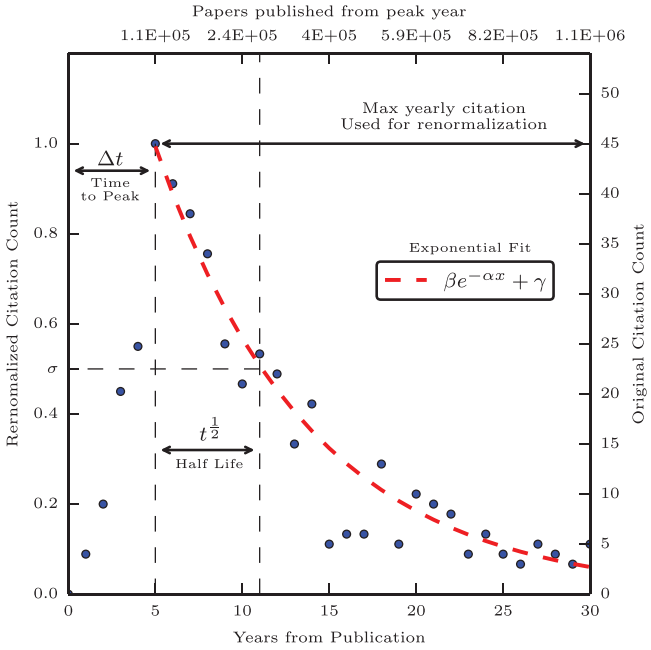

Fig. 2. Schematic representation of the citation evolution of a typical paper.

3.2. Evolution of the time to peak

Next we investigate whether the time to reach the peak in the number of citations Δt_{peak} changes with time. In Fig. 3(a)–(d) we plot the distribution of Δt_{peak} for papers published in the same year, for all four disciplines and for several years. The majority of the papers peak within a few years since publication. Papers in Biology are characterized by small Δt_{peak} as compared to papers in Medicine, Physics and Chemistry. For all fields the distribution of Δt_{peak} is time dependent, with its value decreasing steadily in time. Fig. 3(e) and (f) shows the time evolution of the mean of Δt_{peak} for different fields and

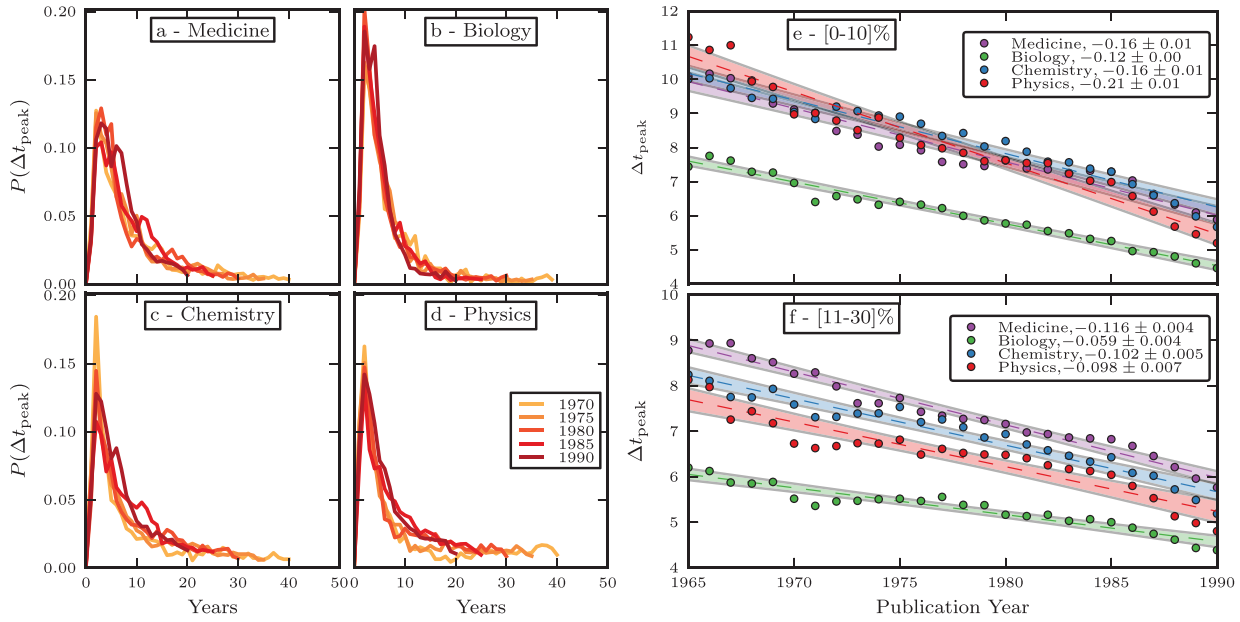

Fig. 3. Time to reach the peak attention Δt_{peak} is both field and time dependent. (a)–(d) Distribution of Δt_{peak} for papers in the top 10% published in the same year, for different fields and publication years. (e) and (f) Time evolution of the mean values of Δt_{peak} for top 10% and [11–30]% percentiles. The mean value $\langle \Delta t_{peak} \rangle$ decreases linearly in time. The linear fit, 95% confidence interval and the slopes of the linear fits are also shown. Papers peaking after 2005 are not considered as their peak years might still be subject to change.

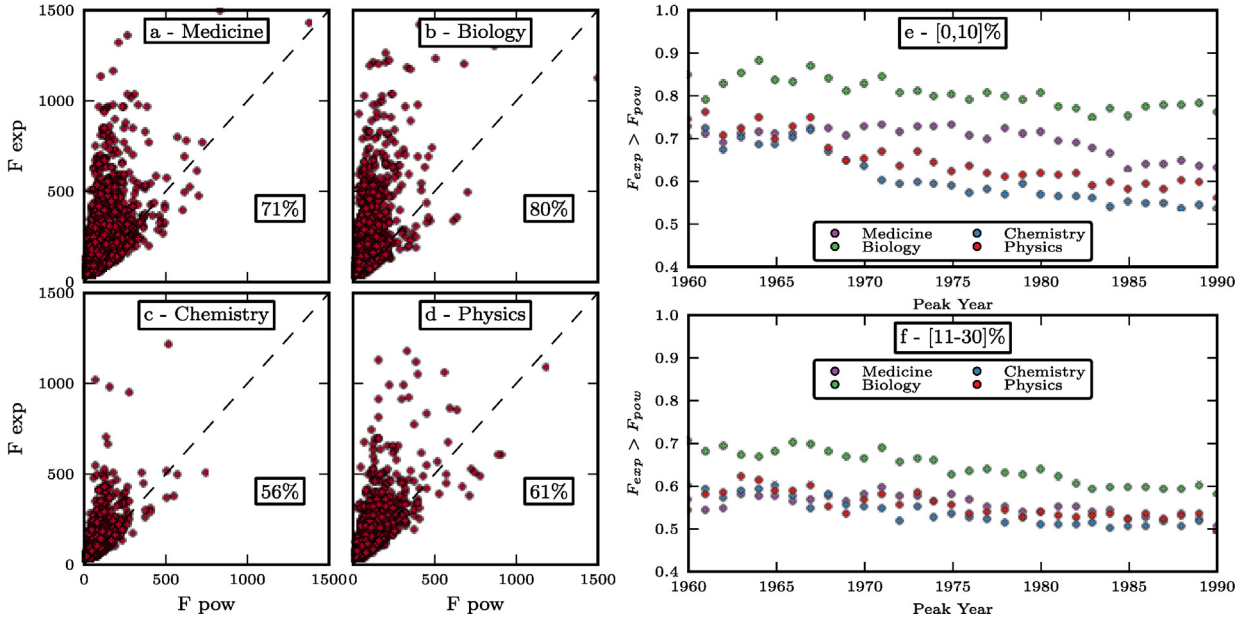

Fig. 4. Comparison of exponential fits with power law fits as described by the *F*-statistics. (a)–(d) Papers peaking in 1980, with the number in the box indicating the percentage of papers better fitted by exponentials than by power laws. In particular, it is worth noticing that there is a significant density of points in the high F_{exp}-low F_{pow} area, showing a series of papers for which the power law fit was clearly outperformed by the exponential fit. There is no trace of the opposite scenario, with papers better fitted by power-law lying close to the diagonal line. (e) and (f) The time evolution of the fraction of papers for which exponentials are better descriptors than power laws, according to the *F*-score, for the top 10% and [11–30]% percentiles papers over different years.

two groups of papers: the most cited 10% and the [11–30] percentile. The decreasing mean of the time to peak indicates that in recent times papers are taking less time to reach the peak of their attention. This result seems to be consistent with previous findings (Egghe, 2010; Larivière, Archambault, & Gingras, 2008) showing, both theoretically and empirically, that the average reference age is an increasing function of time. This would suggest that more recent papers are able to dig deeper in scientific literature, reducing the amount of attention available for papers published in recent years and therefore causing a shortening of the time needed to peak. Also, this behavior is shown to be independent of the citation volume of the papers, although papers with fewer citations take less time to reach the peak. Biology shows again a unique behavior, with its values being constantly below the ones of the other fields, indicating an intrinsic faster peak time.

3.3. Functional form of citation decay

To investigate the time evolution of the change in *attention* we first determine the functional form of the citation decay of each paper. We fit the normalized citation trajectories $\widetilde{c}_i(t) \equiv c_i(t)/c_i^{\max}$ using both the exponential and power law curves. We used an additional parameter in both fitting functions because the normalized citation curves after the initial decay eventually converge to a nonzero plateau. The exponential fitting function is given by $\widetilde{c}_i(t) = \beta_e \exp(-\alpha_e t) + \gamma_e$ whereas the power law fitted function is given by $\widetilde{c}_i(t) = \beta_p t^{-\alpha_p} + \gamma_p$. We fit the normalized citation trajectories of each paper and determine the best fit parameters using the least square method. First, we found that for the majority of the papers both the exponential and power law decrease could fit the decaying behavior, since the *p*-value of the fit is less than 10^{-3}. However, comparing the two fits for each paper using *F*-statistics, we found that the exponential fits better the decaying behavior. Fig. 4 shows that for most paper *F*-statistics is much larger for the exponential fit as compared to the power law fit. Interestingly, in recent years the fraction of papers that fits a power-law curve has been increasing systematically. Fig. 4(e) shows the time evolution of the fraction of papers whose *F*-score in the exponential fitting exceeds the *F*-score for the power law case for the top 10% decile. All the four fields show a trend where the power law fit gradually improves in time. This phenomenon may be linked to the smaller impact of the convergence to the final plateau, on the fit. On average the convergence to the plateau takes more than 20 years, and papers in recent years might not have reached this plateau in their decay.

3.4. Ultradiffusion and decay in attention

A trademark of the evolution of the number of citations of a paper is their decline after reaching a peak. Here, we provide an explanation of this decay. Each citation is considered an *event* and the temporal evolution of the number of citations (after the peak) is taken as a *counting process*. The observed counting process could be rationalized as ultradiffusive if it has signatures associated with an ultradiffusive process. Ultradiffusion is a stochastic process where every timestamp of a

timeseries $\{t_i\}$ ($t_i < t_j$ if $i < j$) $\forall i \in 0 \cdots n$ is associated with an event $\{X_{t_n - t_i}\}$. State $X_{t_n - t_0}$ is analogous to the event of citing the paper. All the other states are associated with not citing the paper. Unlike the Poisson process, which assumes that events occur independently of each other, ultradiffusion elicits that a later event might be caused by or correlated to an earlier event or a combination of earlier events. The earlier event in turn might be independent or might be correlated to a combination of even earlier events. This leads to a hierarchical causal/correlational model of prior event occurrences which can be used to predict the occurrence of a new event. Thus, ultradiffusion proposes that the observed pattern of events is a consequence of an underlying hierarchy of states. In this hierarchical model, an event temporally nearer to the occurring event has a greater probability of affecting it. In other words, the correlation between two events is determined by a notion of "closeness" or distance between them.

For any ultradiffusive process there must be an ultrametric space on which distances between occurrences are defined. In this case the distance between two events X_{t_i} and X_{t_j} can be defined as

$$d(X_{t_i}, X_{t_j}) = \begin{cases} |\max(t_n - t_i, t_n - t_j)|, & \text{if } i \neq j, \\ 0, & \text{otherwise.} \end{cases} \quad (1)$$

The above definition of distance satisfies the *ultrametric distance metric properties* because:

1. $d(X_{t_i}, X_{t_j}) \geq 0$ (non-negative)
2. $d(X_{t_i}, X_{t_j}) = 0$ if $i = j$ (identity of indiscernibles)
3. $d(X_{t_i}, X_{t_j}) = d(X_{t_j}, X_{t_i})$ (symmetry)
4. $d(X_{t_i}, X_{t_j}) \leq \max(d(X_{t_i}, X_{t_k}), d(X_{t_k}, X_{t_j}))$ (ultrametric property).

Therefore the associated space is ultrametric (Ghosh & Huberman, 2014). For an unltradiffusive process, the autocorrelation $P_{X_{t_i}}(t)$, i.e., the probability of finding the system at the initial state X_{t_i} after time t can be calculated analytically. The autocorrelation function has an exact solution for an ultrametric space defined by a hierarchical tree. Assuming that the rate of transition between states is X_{t_i} and X_{t_j} is $e^{-\mu d(X_{t_i}, X_{t_j})}$ and the probability of citing the paper is 1 when the peak in the number of citations is reached, the probability of citing the paper at time t is given by $P_{X_{t_n - t_0}}(t)$. When the number of states is finite, such an autocorrelation function is exponential in nature, otherwise it follows a power law behavior (Bachas & Huberman, 1987).

3.5. Evolution of the decay exponent

Fig. 5 shows the distributions of the exponential decay rates α_e for papers grouped by their peak years. The distributions for different disciplines show that majority of papers have a characteristic rate. Moreover, for all the disciplines the shape of the distribution is broader for papers peaking in recent years. The median of the distributions shows a systematic increase in time (Fig. 5(e) and (f)). Such a faster decay behavior is independent of the fitting ansatz. Furthermore, this pattern is independent of the group of papers chosen for the analysis (top 10% for top panel, [11–30] percentile for bottom panel). This suggests that the later a paper peaks, the shorter is its life cycle, implying a faster decay of scientific attention in terms of absolute time. The decay rates and their relative increase with time appears to be field dependent. For example, for Physics and Chemistry the decay is faster compared with Biology and Medicine.

3.6. Exponential increase in number of publications

The progressively faster decay in attention we observe is compatible with the intuitive picture of scientific theories and papers constantly replaced by other competing results. As the number of publications is also growing with time, it takes less time to replace or update older scientific results. Thus, the rapid increase in the number of papers could provide an explanation. In Fig. 6 we report the growth of the number of publications in different fields with time, fitted by the function $N_p = N_0 \exp^{\delta t}$. All the fields show an exponential increase, as observed for the total number of publications.

Hence, the process of attention gathering needs to take into account the increasing competition between scientific products. With the increase of the number of journals and increasing number of publications in each journal (not to mention the growth of online journals, which do not have physical constraints in their publication volume), a scientist inevitably needs to filter where to allocate its attention, i.e. which papers to cite, among an extremely broad selection. This may also question whether a scientist is actually fully aware of all the relevant results available in scientific archives. Even though this effect is partially compensated by the increase of the average number of references, one needs to consider the impact of increasing publication volume on the attention decay.

3.7. Half-life

To check the robustness of our result that the citation decay rate is becoming faster for recent papers, we measure the *half-life* of each publication. The *half-life* of a paper is a metric regularly adopted to evaluate the typical life-cycle of a paper.

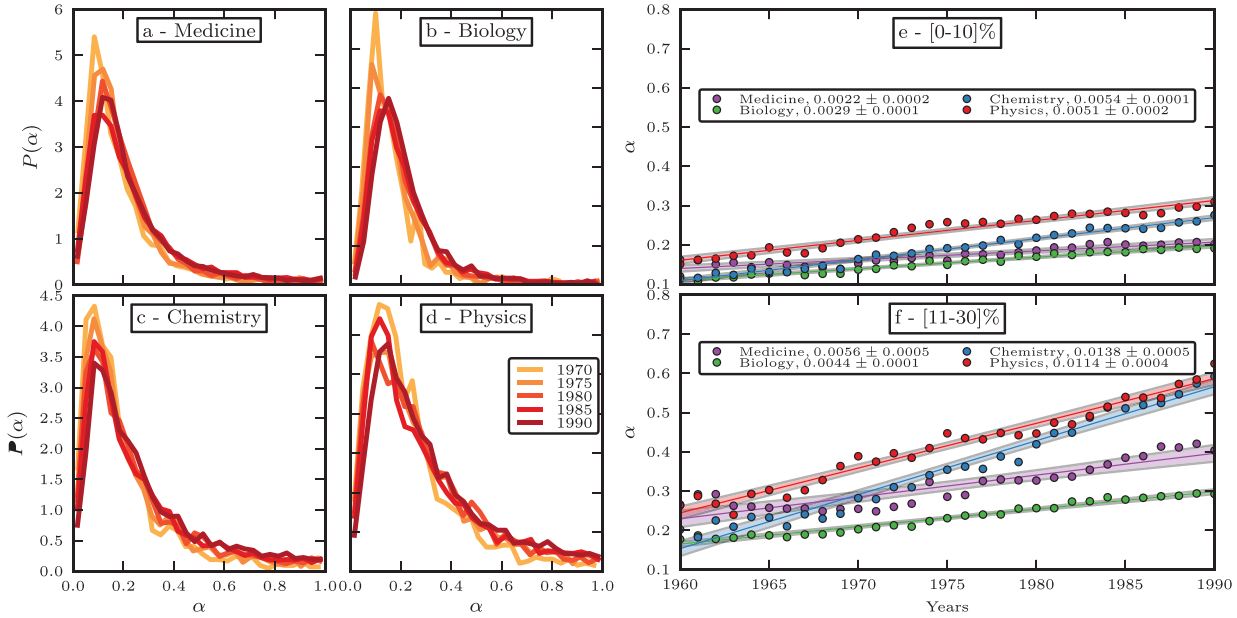

Fig. 5. Attention to publication is decaying faster in time. (a)–(d) Distribution of parameter α for exponential fits in different years for the four disciplines. For recent years the tail of the distribution becomes progressively fatter. (e) and (f) Time evolution of the median of the distributions of the decay rates α, along with linear fit, 95% confidence interval and slopes. The top panel refers to the top 10% most cited papers, the bottom panel to the [11–30] percentile. The data suggests a "grouping" of Medicine and Biology vs Physics and Chemistry, with the two groups having nearly identical numbers for the fit. Moreover, for the [11–30] range the coefficients are nearly doubled compared to [0–10]. This means that the speed of the decay depends on the citation volume of each paper.

The *half-life* of a paper is the time after which the normalized citation rate $\widetilde{c}_i(t)$ is never above 1/2. Similarly, instead of 1/2, other thresholds σ of the citation rate can also be considered. In mathematical terms:

$$t_i^{\frac{1}{2}} = \max\{t \text{s.t.} \widetilde{c}_i(t) \geq \frac{1}{2}\}. \tag{2}$$

The value $t_i^{\frac{1}{2}}$ is the year of the last "sub-peak" of attention for paper i as it quantifies the last moment in the history of the paper at which it has been able to gather sufficient attention. Fig. 7(top panels) shows the time evolution of the

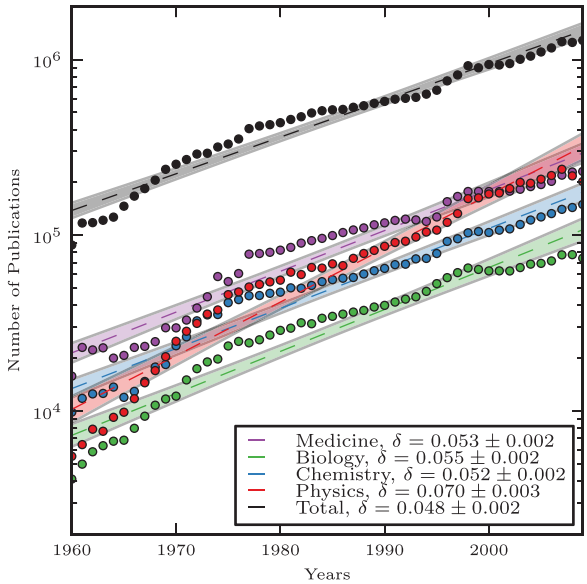

Fig. 6. Increase in the number of publications with time since 1960 along with exponential fits, 95% confidence intervals and rates.

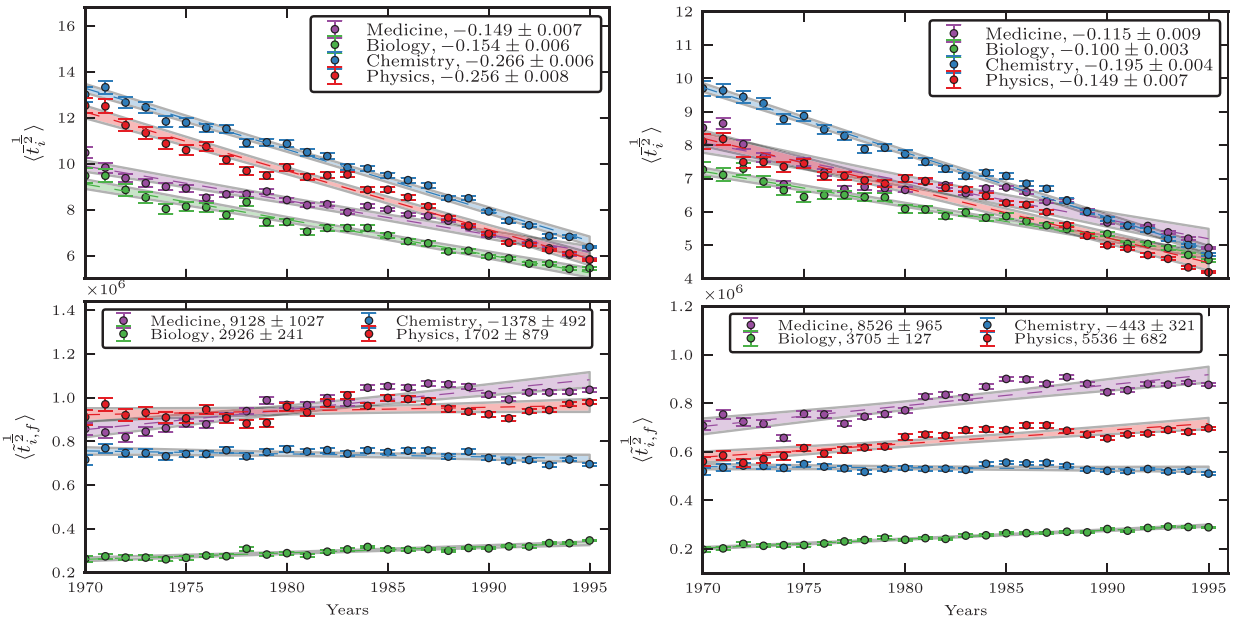

Fig. 7. The half-life of papers $t_i^{\frac{1}{2}}$ in terms of absolute time decreases linearly, whereas the rescaled half-life of papers $\widetilde{t}_i^{\frac{1}{2}}$ in terms of the number of publications is relatively constant. The panels show the evolution of $\langle t_i^{\frac{1}{2}} \rangle$ (top) and $\langle \widetilde{t}_{i,f}^{\frac{1}{2}} \rangle$ (bottom) for the four different fields and for the top 10% (left) and for the [11–30]% percentile (right). The latter values are divided by a large constant to get small values on the *y*-axis, which are easier to display. The error bars indicate the standard errors. Linear fits along with their 95% confidence level intervals are also shown. In the legend the values of the linear coefficients are shown for both absolute (q_y) and renormalized (q_r) time. 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 times.

half-life measure. The mean of the absolute measure $\langle t_i^{\frac{1}{2}} \rangle$ decreases linearly with time for all the four fields. This decrease is consistent with the linear increase in the decay rate of the citation trajectory. Also, there is an interesting grouping between Medicine/Biology and Chemistry/Physics: they start off widely separated but they converge pairwise to similar values in recent years.

3.8. Rescaling time

The *half-life* of a paper can also be used to analyze the impact of the growth of system size. Using the data shown in Fig. 6, we are able to convert its value from a measure of time into a measure of number of publications in the paper's discipline that have been published between the peak of the paper and $t_i^{\frac{1}{2}}$. Therefore we are able to define a renormalized version of $t_i^{\frac{1}{2}}$ as:

$$\widetilde{t}_{i,f}^{\frac{1}{2}} = \sum_{t=t^{\text{peak}}+1}^{t_i^{\frac{1}{2}}} N_p^f(t) \tag{3}$$

where t^{peak} stands for the peak year and $N_p^f(t)$ indicates the number of publications in field *F* of paper *i* for year *t*.

Fig. 7(bottom panels) shows the time evolution of the renormalized half-life measure. Contrary to the previous measure, the evolution of the renormalized half-life $\langle \widetilde{t}_{i,f}^{\frac{1}{2}} \rangle$ shows a relatively stable behavior. Note that, this observation is highly non-trivial as the stable renormalized half-life is only expected in the case when the exponential increase in the number of publications exactly compensate for the decay in citation rate. A similar behavior is also observed when lower thresholds σ are used, i.e., by forcing the drop to be more significant (see Appendix Fig. C.2(a)). The renormalized half life defined in Eq. (3) provides a measure of the time required for a paper to fall below a certain arbitrarily defined threshold of attention in terms of number of publications, which can be seen to represent the amount of "competition" a paper is about to withstand before dropping to significantly lower values of attention.

Interestingly, the picture changes if we consider the half-life to be the first time when the normalized citation rate $\widetilde{c}_i(t)$ decreases below 1/2. In this case, the renormalized half-life shows an increasing pattern with time (Appendix Fig. C.2(b)). Such alternative measure quantifies the time taken to have the first lowest drop of attention. However data suggests that

such value seems to be stable across years for each field as an initial drop in attention appears to be structurally inevitable. This inevitably leads, after renormalization, to a significantly increasing behaviour.

Fig. 7 suggests that, even though papers are now taking on average less time to drop below a certain threshold of attention, the number of published papers after which a work becomes obsolete does not show the same behavior. On the contrary, our data indicates an approximately constant value throughout the time period of the study. So, the growing number of publications proportionally increases the likelihood of a paper to become obsolete, but the contribution of each paper to this process is about the same, regardless of the age of the paper.

4. Conclusions

We have studied how attention towards scientific publications diminishes over time, due to the obsolescence of knowledge. For millions of papers in four different disciplines we find that after reaching a peak, typically a few years since publication, the number of citations goes down relatively fast. We find that exponential decays are to be generally preferred over power law decays, though the latter are providing better and better descriptions of the data for recent times. The existence of many time-scales in citation decay and our ability to construct an ultrametric space to represent this decay, leads us to speculate that citation decay is an ultradiffusive process, like the decay of popularity of online content. Interestingly, the decay is getting faster and faster, indicating that scholars "forget" more easily papers now than in the past. We found that this has to do with the exponential growth in the number of publications, which inevitably accelerates the turnover of papers, due to the finite capacity of scholars to keep track of the scientific literature. Although search engines and digitalization have made it easier for scientists to discover relevant information, the amount of information that can be successfully processed is still limited. In fact, by measuring time in terms of the number of published works, the decay appears approximately stable over time, across disciplines, although there are slight monotonic trends for Medicine and Biology. However, we must emphasise that we normalized time by using the number of published papers in the discipline at study. This is the simplest choice to make, but it is not necessarily the most sensible one. The fields we considered are rather broad, and subdivided in many different topics. Scholars working on any of such topics will be affected mostly by the literature of the topic, and hardly by anything else. It is very difficult to isolate the relevant literature case by case. Still, considering the whole bulk of publications in each single discipline is a way to discount the exponential growth of scientific output and we have found that this suffices to counterbalance (at least to a large extent) the apparent faster decay of attention observed in recent years.

Author contributions

All authors designed the research and participated in the writing of the manuscript. PDBP and RKP collected and analysed the data. PDBP performed the research.

Acknowledgements

We used data from the Science Citation Index Expanded, Social Science Citation Index and Arts & Humanities Citation Index, prepared by Thomson Reuters, Philadelphia, Pennsylvania, USA, Copyright Thomson Reuters, 20. We gratefully acknowledge KNOWeSCAPE, COST Action TD1210 of the European Commission, for fostering interactions with leading experts in science of science who gave feedback on the paper. We also thank HP Labs for supporting the visit of SF, during which the project was started.

Appendix A. Description of the categories

To categorize each paper according to its field of publication we use the Thomson Reuters (TR) subject categories. We then aggregated these subject categories into broader scientific fields. A detailed description is provided in Table A.1

Table A.1
Aggregation of TR subject categories in broader fields.

Fields	TR subject categories
Physics	IMAGING SCIENCE &PHOTOGRAPHIC TECHNOLOGY; PHYSICS, APPLIED; OPTICS; INSTRUMENTS &INSTRUMENTATION; PHYSICS, CONDENSED MATTER; PHYSICS, FLUIDS &PLASMAS; PHOTOGRAPHIC TECHNOLOGY; PHYSICS, ATOMIC, MOLECULAR &CHEMICAL; ACOUSTICS; PHYSICS; PHYSICS, MATHEMATICAL; MECHANICS; PHYSICS, NUCLEAR; SPECTROSCOPY; THERMODYNAMICS; PHYSICS, PARTICLES &FIELDS; NUCLEAR SCIENCE &TECHNOLOGY; PHYSICS, MULTIDISCIPLINARY; ASTRONOMY &ASTROPHYSICS;
Chemistry	CHEMISTRY, INORGANIC &NUCLEAR; ELECTROCHEMISTRY; CHEMISTRY, PHYSICAL; CHEMISTRY, ANALYTICAL; POLYMER SCIENCE; CHEMISTRY, MULTIDISCIPLINARY; CRYSTALLOGRAPHY; CHEMISTRY, APPLIED; CHEMISTRY; CHEMISTRY, ORGANIC;
Molecular Biology	BIOCHEMICAL RESEARCH METHODS; BIOCHEMISTRY &MOLECULAR BIOLOGY; BIOMETHODS; BIOPHYSICS; CELL &TISSUE ENGINEERING; CELL BIOLOGY; CYTOLOGY &HISTOLOGY; MATHEMATICAL &COMPUTATIONAL BIOLOGY; MICROSCOPY;

Table A.1 (*Continued*)

Fields	TR subject categories
Physiology or Medicine	CYTOLOGY &HISTOLOGY; BIOCHEMISTRY &MOLECULAR BIOLOGY; CELL BIOLOGY; BIOCHEMICAL RESEARCH METHODS; CELL &TISSUE ENGINEERING; MATHEMATICAL &COMPUTATIONAL BIOLOGY; BIOPHYSICS; BIOMETHODS; MICROSCOPY; ENGINEERING, BIOMEDICAL; IMMUNOLOGY; MEDICAL LABORATORY TECHNOLOGY; MEDICINE, RESEARCH &EXPERIMENTAL; PARASITOLOGY; PHYSIOLOGY; ANATOMY &MORPHOLOGY; PATHOLOGY; ONCOLOGY; RHEUMATOLOGY; VASCULAR DISEASES; PSYCHIATRY; GERIATRICS &GERONTOLOGY; DENTISTRY, ORAL SURGERY &MEDICINE; OPHTHALMOLOGY; DENTISTRY ORAL SURGERY &MEDICINE; MEDICINE, LEGAL; EMERGENCY MEDICINE &CRITICAL CARE; CLINICAL NEUROLOGY; TRANSPLANTATION; HEMATOLOGY; INFECTIOUS DISEASES; RESPIRATORY SYSTEM; PERIPHERAL VASCULAR DISEASE; MEDICINE, GENERAL &INTERNAL; PEDIATRICS; EMERGENCY MEDICINE; INTEGRATIVE &COMPLEMENTARY MEDICINE; GASTROENTEROLOGY &HEPATOLOGY; DERMATOLOGY; REHABILITATION; ANESTHESIOLOGY; TROPICAL MEDICINE; MEDICINE, MISCELLANEOUS; ENDOCRINOLOGY &METABOLISM; NEUROIMAGING; ANDROLOGY; ORTHOPEDICS; OBSTETRICS &GYNECOLOGY; ALLERGY; CRITICAL CARE MEDICINE; OTORHINOLARYNGOLOGY; RADIOLOGY, NUCLEAR MEDICINE &MEDICAL IMAGING; SURGERY; CARDIAC &CARDIOVASCULAR SYSTEMS; DERMATOLOGY &VENEREAL DISEASES; AUDIOLOGY &SPEECH-LANGUAGE PATHOLOGY; RADIOLOGY &NUCLEAR MEDICINE; UROLOGY &NEPHROLOGY; CRITICAL CARE; CARDIOVASCULAR SYSTEM;

Appendix B. Evolution of the number of citations for other decile

Fig. B.1 is the analog of figure Fig. 1 of the main text, but is focused on the top [11–30]% papers (based on their total number of citations). Compared to the original figure the values of $\langle \tilde{c}(t) \rangle$ is lower, linked to the fact that these papers have accumulated fewer citations. The top panels (A,B), where the papers are grouped by their publication year, show that the average peak is more concentrated in the initial years and is followed by a more rapid decay. Finally, the citation trajectories reach a plateau that is significantly lower than the respective one for the top decile. Similarly, the papers grouped by their

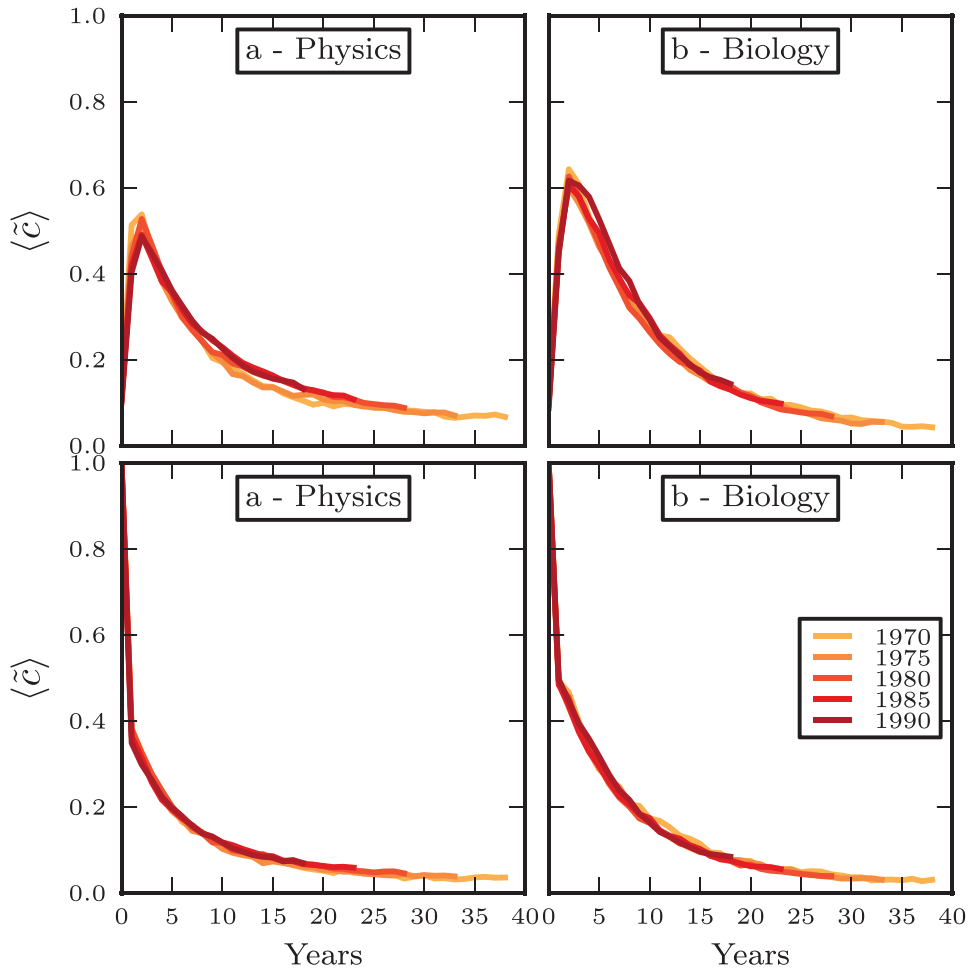

Fig. B.1. Averaged citation trajectories are calculated for papers in the [11–30]% window based on their total number of citations.

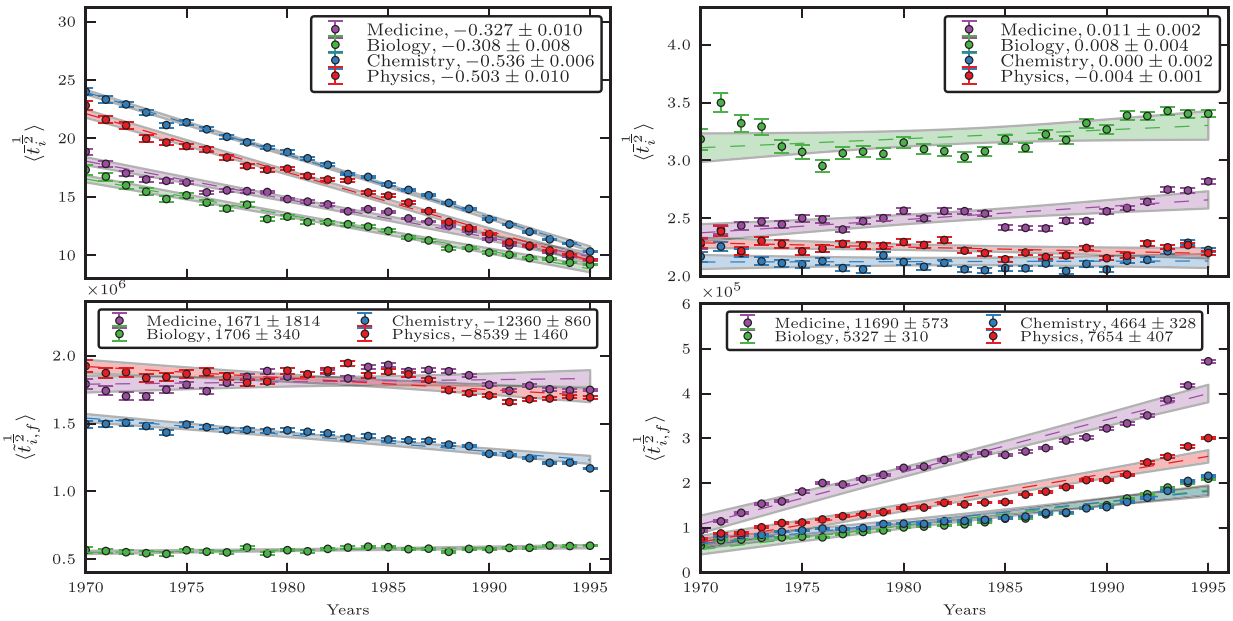

Fig. C.2. (Left) The half-life of papers $t_i^{\frac{1}{2}}$ with $\sigma = 0.3$. (Right) The alternative half-life of papers $\bar{t}_i$ with $\sigma = 0.5$.

peak year (bottom panels, C,D), also show a larger drop in $\langle \widetilde{c}(t) \rangle$ in the first few years followed by a lower value of the final plateau.

Appendix C. Evolution of half-life for different values of σ and alternative definition of half-life

Fig. C.2(a) and (b) is the analogous of Fig. 7 with $\sigma = 0.3$. This implies choosing a lower threshold for the definition of the point below which a paper is considered to have completed its life cycle. Data suggests that the pattern shown in the paper is retained for other choices of parameters. However, at $\sigma = 0.3$, Physics also shows a slight decreasing patter, whereas Medicine and Biology retain their increasing trends.

Fig. C.2(c) and (d) is the analog of the previous figure, with the alternative half-life defined as

$$\bar{t}_i^{\frac{1}{2}} = \min\{t \text{s.t.} \widetilde{c}_i(t) \leq \frac{1}{2}\}. \tag{C.1}$$

whereas $\bar{t}$ is defined still in the same way as in Eq. (3) but using the previously defined value for t. In th framework the half life of the paper is considered as the first year in its life cycle where its citations have dropped below a certain threshold. The figure shows that with this definition the values of $\bar{t}$ lose their decreasing pattern in favour of a field specific value, which is retained in the years. Similarly, the behavior for $\bar{t}$ shows a deviation from the previously constant pattern in favor of a significant increase in its values.

References

Avramescu, A. (1979). Actuality and obsolescence of scientific literature. *Journal of the American Society for Information Science*, *30*(5), 296–303.
Bachas, C. P., & Huberman, B. A. (1987). Complexity and ultradiffusion. *Journal of Physics A: Mathematical and General*, *20*(14), 4995, http://stacks.iop.org/0305-4470/20/i=14/a=036.
Bouabid, H. (2011). Revisiting citation aging: A model for citation distribution and life-cycle prediction. *Scientometrics*, *88*(1), 199–211.
Bouabid, H., & Larivière, V. (2013). The lengthening of papers life expectancy: A diachronous analysis. *Scientometrics*, *97*(3), 695–717.
Crane, R., & Sornette, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proceedings of the National Academy of Sciences*, *105*(41), 15649–15653.
Dezsö, Z., Almaas, E., Lukács, A., Rácz, B., Szakadát, I., & Barabási, A.-L. (2006). Dynamics of information access on the web. *Physical Review E*, *73*(6), 066132.
Dukas, R. (2004). Causes and consequences of limited attention. *Brain, Behavior and Evolution*, *63*, 197–210.
Egghe, L. (2000). Aging, obsolescence, impact, growth, and utilization: Definitions and relations. *Journal of the American Society for Information Science and Technology*, *51*(11), 1004–1017.
Egghe, L. (2010). A model showing the increase in time of the average and median reference age and the decrease in time of the Price Index. *Scientometrics*, *82*(2), 243–248.
Franck, G. (1999). Scientific communication – A vanity fair? *Science*, *286*(5437), 53–55. http://www.sciencemag.org/content/286/5437/53
Ghosh, R., & Huberman, B. (2014). Information relaxation is ultradiffusive. In *Proceedings of 2014 ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conf* 2014. ASE.
Huberman, B. A., Romero, D. M., & Wu, F. (2009). Crowdsourcing attention and productivity. *Journal of Information Science*, *35*(6), 758–765, http://jis.sagepub.com/content/35/6/758.
Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.

Klamer, A., & Dalen, H. P. V. (2002). Attention and the art of scientific publishing. *Journal of Economic Methodology*, *9*(3), 289–315. http://dx.doi.org/10.1080/1350178022000015104
Larivière, V., Archambault, E., & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900-2004). *Journal of the American Society for Information Science and Technology*, *59*(2), 288–296.
Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). 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* USA, (pp. 497–506). New York, NY: ACM.
Matsubara, Y., Sakurai, Y., Prakash, B. A., Li, L., & Faloutsos, C. (2012). Rise and fall patterns of information diffusion: Model and implications. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 6–14). ACM.
Medo, M. c. v., Cimini, G., & Gualdi, S. (2011). Temporal effects in the growth of networks. *Physical Review Letters*, *107*, 238701. http://dx.doi.org/10.1103/PhysRevLett.107.238701
Nakamoto, H. (1988). Synchronous and dyachronous citation distributions. In L. Egghe, & R. Rousseau (Eds.), *Informetrics 87/88* (pp. 157–163). Amsterdam: Elsevier Science Publisher.
Pashler, H. E. (1998). *The psychology of attention*. Cambridge, MA: MIT Press.
Petersen, A. M., Fortunato, S., Pan, R. K., Kaski, K., Penner, O., Rungi, A., Riccaboni, M., Stanley, H. E., & Pammolli, F. (2014). Reputation and impact in academic careers. *Proceedings of the National Academy of Sciences*, *111*(43), 15316–15321. http://www.pnas.org/content/111/43/15316
Pieters, R., Rosbergen, E., & Wedel, M. (1999). Visual attention to repeated print advertising: A test of scanpath theory. *Journal of Marketing Research*, 424–438.
Pollman, T. (2000). Forgetting and the ageing of scientific publications. *Scientometrics*, *47*(1), 43–54.
Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today*, *58*(6), 49–54.
Reis, R. (2006). Inattentive consumers. *Journal of Monetary Economics*, *53*(8), 1761–1800.
Weng, L., Flammini, A., Vespignani, A., & Menczer, F. (2012). Competition among memes in a world with limited attention. *Sci. Rep.*, 2.
Wu, F., & Huberman, B. A. (2007). Novelty and collective attention. *Proceedings of the National Academy of Sciences*, *104*(45), 17599–17601. http://www.pnas.org/content/104/45/17599.short
Yang, S., Ma, F., Song, Y., & Qiu, J. (2010). A longitudinal analysis of citation distribution breadth for Chinese scholars. *Scientometrics*, *85*(3), 755–765.

Publication III

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

Uncovering the Dynamics of Ego Networks of Scientific Gems

Pietro Della Briotta Parolo[a], Santo Fortunato[a]

[a] *Complex Systems Unit, Aalto University School of Science, P.O. Box 12200, FI-00076, Finland*

Abstract

A promising way to keep track of the impact of a scientific work is to investigate the structure of its ego-network, i. e., of the network consisting of all papers citing that work and their mutual citations. Here we study the dynamics of ego-networks of highly cited articles, and find that it has some peculiar general features. Partial ego-networks, whose papers are published within a sliding time window, are usually very compact in the first years after the publication of the ego-paper, while they eventually fragment in many disconnected components. Their average size peaks after 6-7 years since the publication of the ego-paper and then it steadily decays. These results indicate that in most cases a highly cited paper starts losing visibility within a few years from publication, and its impact is reflected in the success of followup works. The fragmentation of the ego-networks may be due to an increased specialisation of the field of the ego-paper, or a growing popularity of the paper in different fields.

Keywords: Ego Networks, Citation count, Scientometrics

1. Introduction

In social network analysis an ego-network (EN), or ego-centered network, is the graph formed by the neighbours of a specific individual (the ego) and by their mutual relationships [1, 2, 3, 4, 5, 6, 7, 8].

The people in the EN are the ones having the greatest impact on the life of the ego, influencing his attitudes, norms, values, goals and perceptions of the world. Moreover, they are the ones to whom the ego must turn to seek information, help and support. ENs are thus a useful tool to look at social networks from a local perspective.

The concept can be exported to other contexts. For instance, the EN of a scientific paper is the set of all papers citing it, along with their mutual citations. Just like social ENs allow us to uncover the social world of single persons, we can use citation ENs to investigate the impact dynamics of a paper, which is the goal of this work.

We consider ENs of highly cited papers, and study how their properties change in time. To study the dynamics we focus on subsets of each EN, consisting of all papers published in sliding time windows, and their mutual citations. We also investigate the evolution of the full network.

2. Material and methods

2.1. Data description

Our data set consists of all publications (articles and reviews) written in English till the end of 2013 included in the database of Thomson Reuters (TR) Web of Science. We selected recent, highly cited papers, i. e., published after 1990. In the main text we will focus on papers with at least 1000 citations in the first ten years, which add up to more than 2000 papers. In the Appendix we shall vary the threshold to check the robustness of the observed patterns. For each paper, we built the EN of the original work, by selecting the citing papers and returning the citations between them. The networks are created in windows of size w, which indicate how many consecutive years of scientific publications are considered in the analysis.

Fig.1 shows the EN for the famous paper by Barabási and Albert on preferential attachment in complex networks [9]. All the nodes represent papers citing it, along with the connections (citations) between them. A

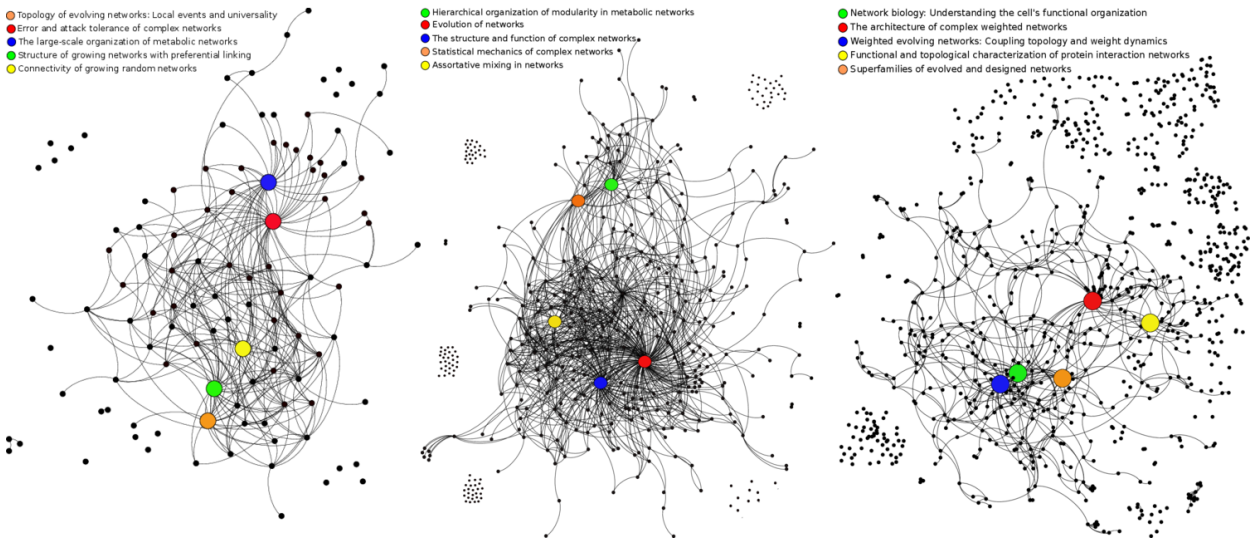

Figure 1: Ego-network for the paper *Emergence of Scaling in Random Networks* (Science 286, 509-512, 1999), by A.-L. Barabási and R. Albert. We consider windows of size $w = 2$ at t=1 (left), t=3 (center) and t=5 (right), where t is the number of years from publication. Therefore the windows are non-overlapping 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.

quick analysis shows some general features of EN evolution: in the beginning they are very dense (left figure), but with the emergence of isolated nodes after a few years (central figure); as time goes by connectivity decreases dramatically and most of the network is made up of isolated nodes and a relatively small connected core.

3. Results and discussions

Properties of the Network

Fig. 2 shows the distribution of EN size for two different window sizes, w=2 (top) and w=3 (bottom) and at different stages in time. The early distribution for the first complete window (the year of publication is not part of it) shows a peak at around 200 ($w = 2$) or 300 ($w = 3$), indicated by the dark red line. Then in the following years the peak moves right (coherently with an increase in citation volume), followed by a retreat until in about 10/12 years the distribution looks similar to the earliest ones. After that, the peak keeps shifting to the left, indicating a decrease in citation volume and a decay of the attention towards the ego-paper [10, 11, 12, 13]. The width of the distributions instead increases with time.

The time evolution of the average EN size can be seen in Fig. 3, for the partial ENs [left panels: w=2 (top), w=3 (bottom)] and for the full one (right panel). For the partial ENs there is an early increase in the mean and median values, followed by a rapid decrease after a broad peak at around $t = 7$. The median falls faster than the mean, suggesting that many networks become small, while a few remain still large. On the other hand the size of the full EN can only increase. The figure shows that the growth is roughly linear in time.

Next, we focus on the structure of the ENs. The first question is whether the network is connected or fragmented into smaller pieces. Figs. 4 and 5 attempt at providing an answer by showing the distributions at different stages of the relative size of the largest connected component (LCC), along with their time evolution. For the partial ENs, the relative size of LCC has initially a rather flat distribution (Fig. 4), so there is a broad range of values that the initial peak can reach. Then the distributions become more and more peaked and shifted to the left, indicating a shrinking of the LCC. The time evolution of the averages can be seen in Fig. 5. In this case there is a more stable pattern, with values starting off around 0.5 and

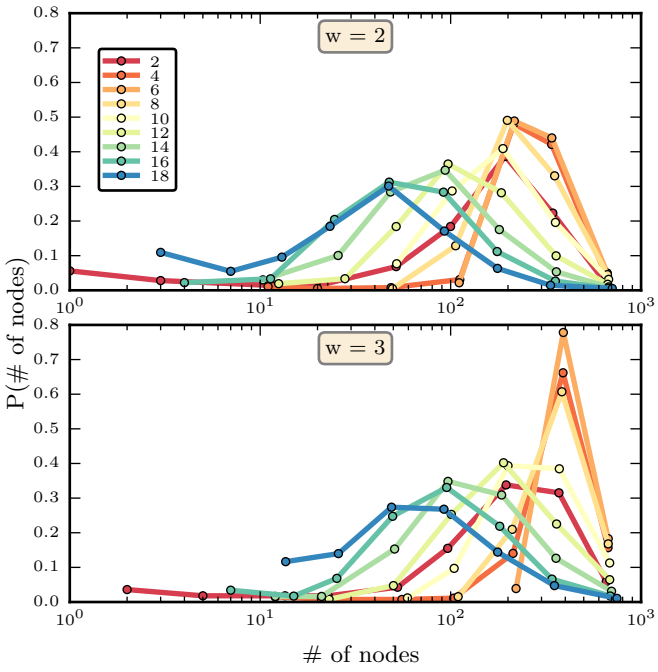

Figure 2: Distribution of EN size, with windows of 2 years (left) and 3 years (right). The different lines indicate different intervals t from the publication of the ego.

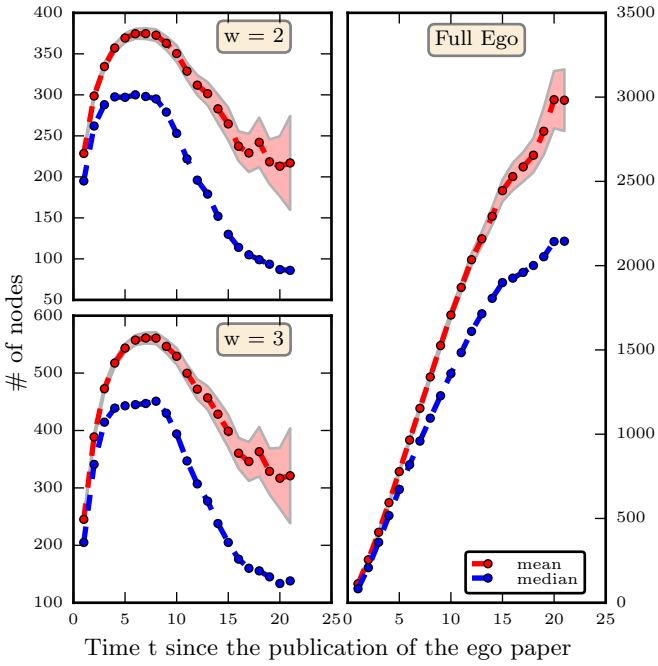

Figure 3: Evolution of mean and median absolute size of the ENs. We also show the standard deviation of the sampled mean. The panels correspond to $w = 2$ (top left), $w = 3$ (bottom left) and to the full EN (right).

then rapidly falling to a plateau which depends on w. For the full EN, as expected, the LCC grows steadily in its initial years, stabilizing at high values after 5 to 10 years from publication. If put together with Fig. 3 this shows that, after the initial transient, papers cite papers in the LCC at an approximately constant rate and the overall connectivity remains strong. Hence we see for the initial years ($2 \leq t \leq 10$) two contrasting

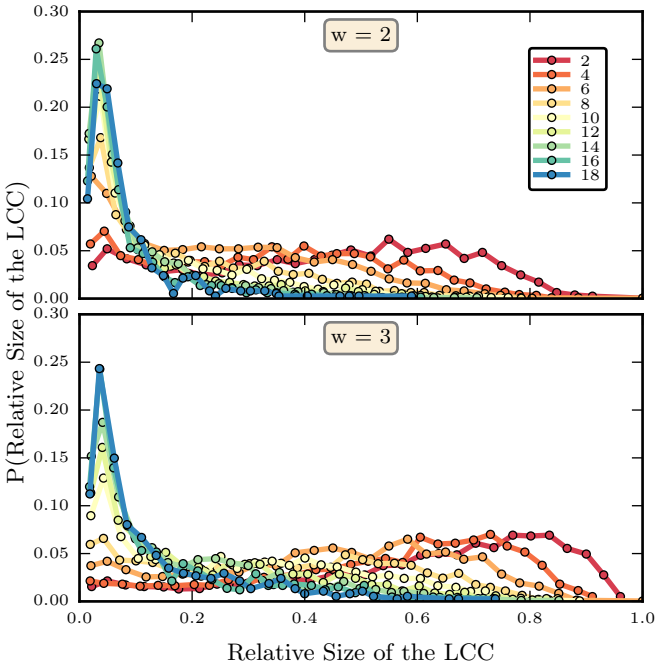

Figure 4: Distribution of the relative fraction of the largest connected component of the EN for all papers with windows of 2 years (top) and 3 years (bottom). The different lines indicate different distances t from publication.

phenomena for the partial ENs: an expansion and subsequent contraction of the size of the network and a constant contraction of the relative size of the LCC. What about the *absolute* size of the LCC? Fig. 6 shows the time evolution of the absolute size of the LCC. We can see a very interesting pattern here. After a few initial years of increase (more required for larger w) the mean absolute size of the LCC falls exponentially until $t \approx 15$, where the data starts becoming noisy due to the lower number of papers citing the ego. The result indicates that, as time goes by, recent papers are less likely to cite each other. Further analysis suggests (Figs. A1 and A2) that the collapse of the largest connected component in the window scenario is not associated to fragmentation, but rather to the disintegration of the network. Even though the relative size of the second largest component grows, its absolute size shows a small increase within the first 10 years, followed by a decreasing trend. The full EN consists almost entirely of the LCC, the rest of the nodes forming very small components. Also, Fig A5 shows the impact of the number of citations of the ego on the exponential decay.

Furthermore, one can look at properties of individual nodes (see also Appendix Figs. A3 and A4) that further characterize the disintegration of the network and of the LCC. Fig. 7 shows the time evolution of the fraction of nodes with incoming degree $k_{in} > 0$, i.e. receiving at least one citation. Consistently with what we have seen before, less and less papers manage to gather any citation within the chosen time window. New papers tend more likely to attach, if at all, to older papers. This can be seen in the full EN, where the fraction of papers receiving citations saturates initially, but keeps slightly increasing. This means that papers keep receiving citations from the other papers of the EN, but not from those within the same time window. Similarly, Fig. 8 shows the time evolution of the fraction of nodes with outgoing degree $k_{out} = 0$, i. e. citing no paper of the EN. This conferms what could be suggested from the previous figure, as on one hand we see that a high fraction of nodes does not contribute citations within their time window, while the total EN keeps receiving citations from the new nodes.

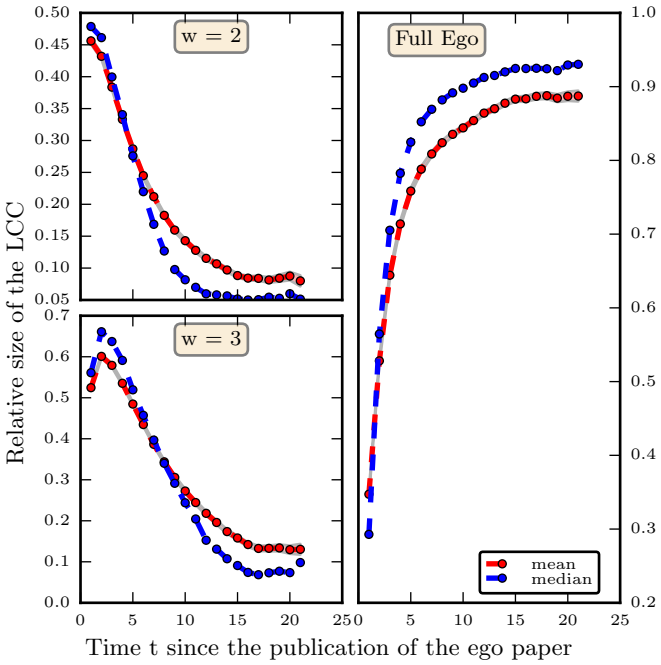

Figure 5: Time evolution of mean and median relative size of the LCC along with the standard deviation of the sampled mean for $w = 2$ (top left) and $w = 3$ (bottom left) and of the full EN (right).

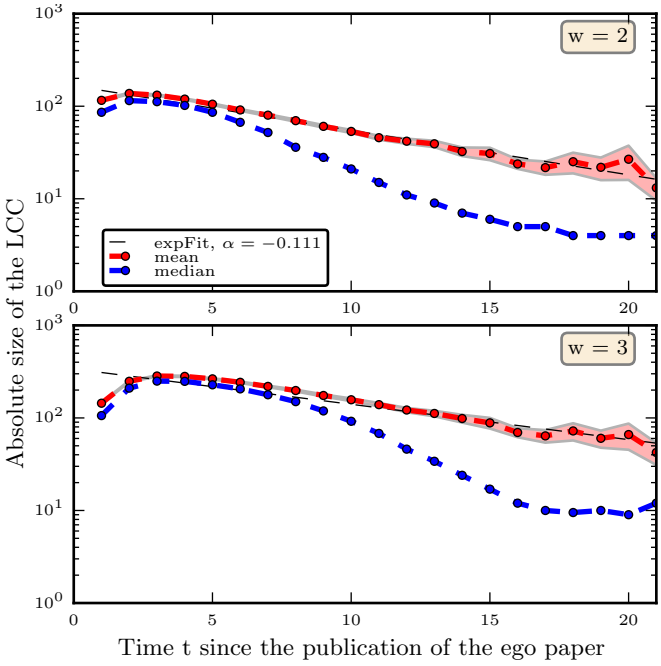

Figure 6: Time evolution of mean and median absolute size of the LCC for $w = 2$ (top) and $w = 3$ (bottom) along with the best fit to the exponential $t = \beta \exp(-\alpha t)$.

Finally, one can look at the relationship between the EN and the overall citation network in order to analyze the impact of the ego paper within its "community". When a new layer of nodes is introduced, each node provides d_i new links. These are just a fraction of the total number of references r_i of the paper. Hence, we can calculate for each paper the value of the fraction $f_i = \frac{d_i}{r_i}$, which quantifies what portion of the reference list goes to members of the EN. Fig. 9 shows the time evolution of the average of this property. As we can see the number initially increases, reaching the peak around 6/7 years after publication, then it decreases. Fig. A6 in the Appendix shows how the number of citations of the ego affects the shape of the curve.

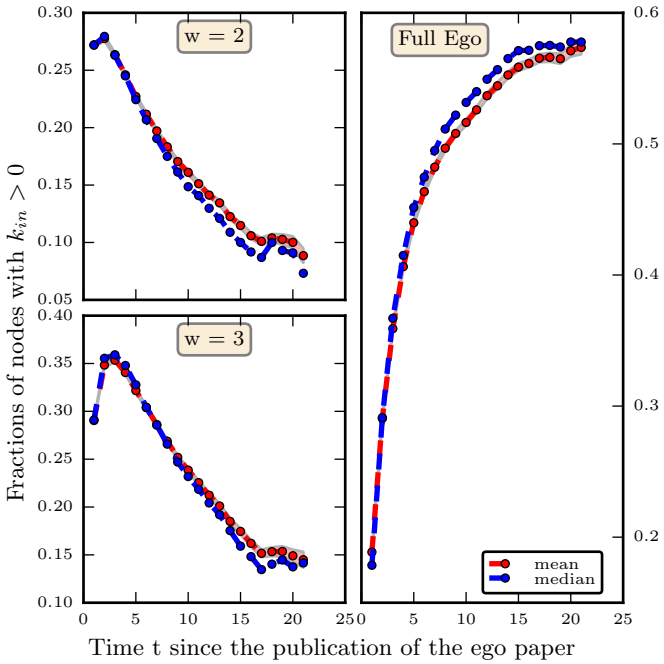

Figure 7: Time evolution of mean and median fraction of nodes with at least one incoming connection from the other nodes of the EN ($k_{in} > 0$) for $w = 2$ (top left), $w = 3$ (bottom left) and for the full EN (right).

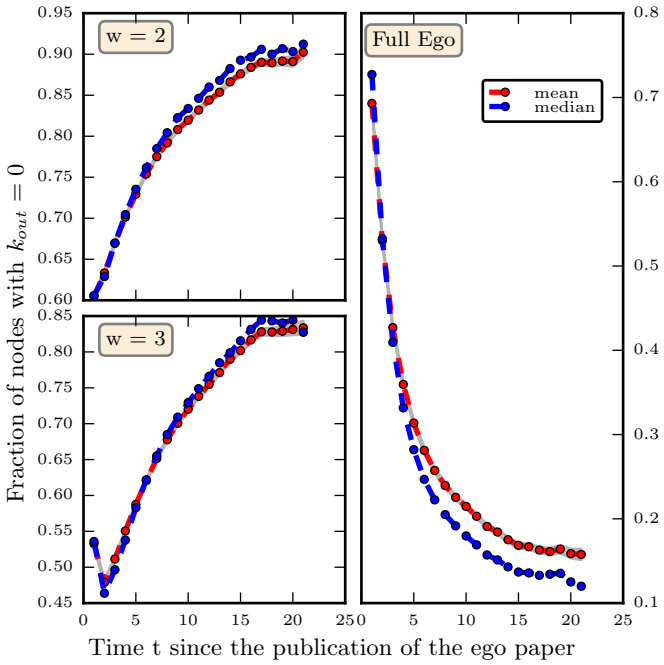

Figure 8: Time evolution of mean and median fraction of nodes without outgoing connections to the other nodes of the EN ($k_{out} = 0$) for $w = 2$ (top left), $w = 3$ (bottom left) and for the full EN (right).

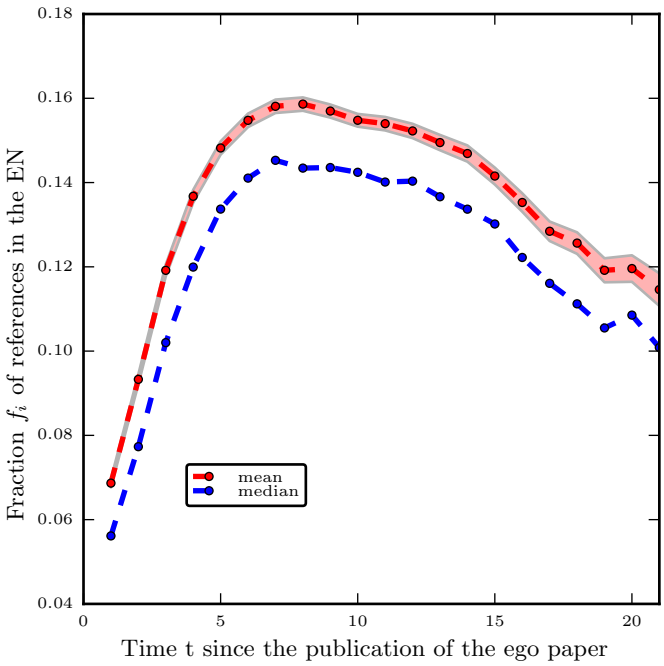

Figure 9: Time evolution of the mean and median of the fraction f_i of references of papers of the full EN belonging to the EN as a function of the number of years since publication.

4. Conclusions

Ego-networks of papers could help us investigate the impact that scientific works have in the literature. We focused on partial ENs, comprising papers published in sliding time windows. We find a consistent scenario, in which the networks fragment into many small components few years after the publication of the ego-paper. The progressive decrease of citations between later members of the EN may signal a specialization of the topic and (or) an increasing popularity of the ego in different disciplines, where citations are infrequent between works on different subjects.

A natural next step of this investigation is proposing a model that describes and possibly predicts the evolution of ENs. Candidate models could build upon popular models of growth of citation networks [14, 15].

Acknowledgments

We used data from the Science Citation Index Expanded, Social Science Citation Index and Arts & Humanities Citation Index, prepared by Thomson Reuters, Philadelphia, Pennsylvania, USA, Copyright Thomson Reuters, 2013. This research is supported by the European Community's H2020 Program under the scheme ÍNFRAIA-1-2014-2015: Research Infrastructures', grant agreement 654024 *SoBigData: Social Mining & Big Data Ecosystem*, http://www.sobigdata.eu.

Author Contributions

Both authors designed the research and participated in the writing of the manuscript.

References

[1] E. Bott, Family and social network: Roles, norms and external relationships in ordinary urban families, Tavistock Publications, 1957.
[2] L. C. Freeman, Centered graphs and the structure of ego networks, Mathematical Social Sciences 3 (3) (1982) 291–304.
[3] P. D. Killworth, E. C. Johnsen, H. R. Bernard, G. A. Shelley, C. McCarty, Estimating the size of personal networks, Social Networks 12 (4) (1990) 289–312.
[4] S. Wasserman, K. Faust, Social network analysis: Methods and applications, Vol. 8, Cambridge university press, 1994.
[5] M. E. Newman, Ego-centered networks and the ripple effect, Social Networks 25 (1) (2003) 83–95.
[6] J. Scott, Social network analysis, Sage, 2012.
[7] V. Arnaboldi, M. Conti, A. Passarella, 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), IEEE, 2012, pp. 31–40.
[8] J. J. McAuley, J. Leskovec, Learning to discover social circles in ego networks., in: NIPS, Vol. 2012, 2012, pp. 548–56.
[9] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512.
[10] A. Avramescu, Actuality and obsolescence of scientific literature, Journal of the American Society for Information Science 30 (5) (1979) 296–303.
[11] T. Pollman, Forgetting and the ageing of scientific publications, Scientometrics 47 (1) (2000) 43–54.
[12] H. Bouabid, V. Larivière, The lengthening of papers life expectancy: a diachronous analysis, Scientometrics 97 (3) (2013) 695–717.
[13] P. D. B. Parolo, R. K. Pan, R. Ghosh, B. A. Huberman, K. Kaski, S. Fortunato, Attention decay in science, Journal of Informetrics 9 (4) (2015) 734–745.
[14] Y.-H. Eom, S. Fortunato, Characterizing and modeling citation dynamics, PloS one 6 (9) (2011) e24926.
[15] D. Wang, C. Song, A.-L. Barabási, Quantifying long-term scientific impact, Science 342 (6154) (2013) 127–132.

Appendix

Further network properties.

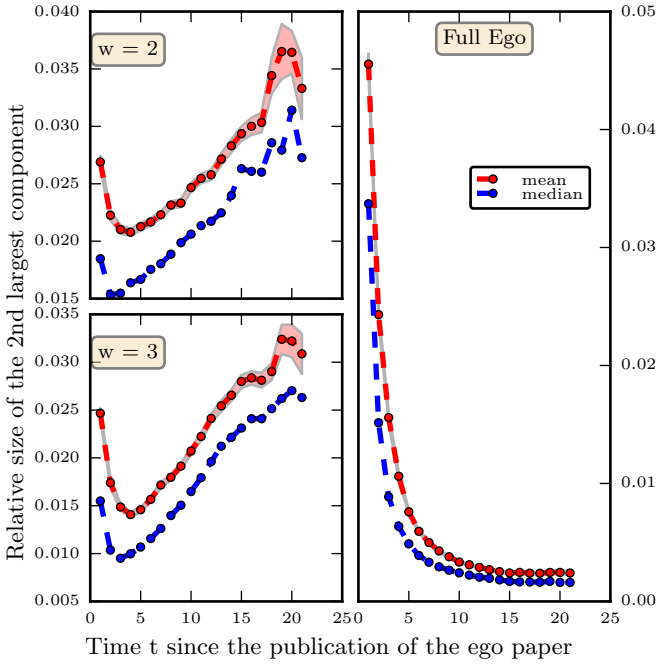

Figure A1: Time Evolution of mean and median relative size of the second largest component for $W = 2$ (top) and $W = 3$ (bottom) (left) and of the full EN (right).

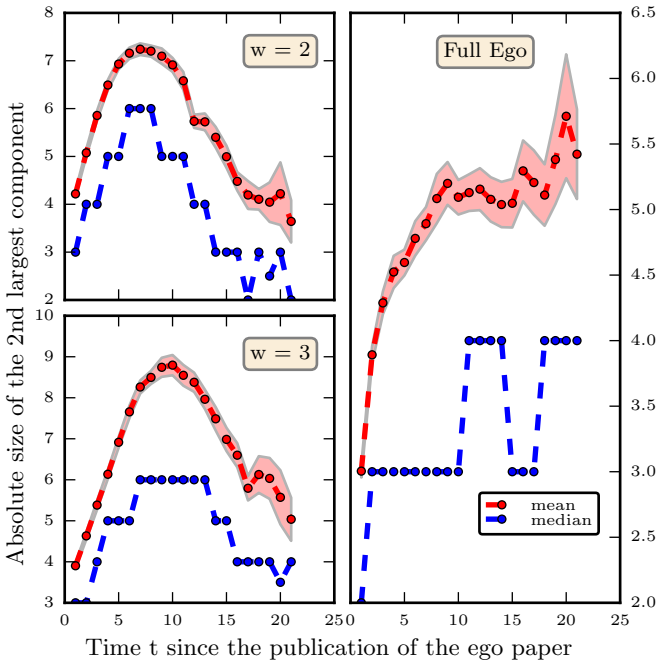

Figure A2: Time Evolution of mean and median absolute size of the second largest component for $W = 2$ (top) and $W = 3$ (bottom) (left) and of the full EN (right).

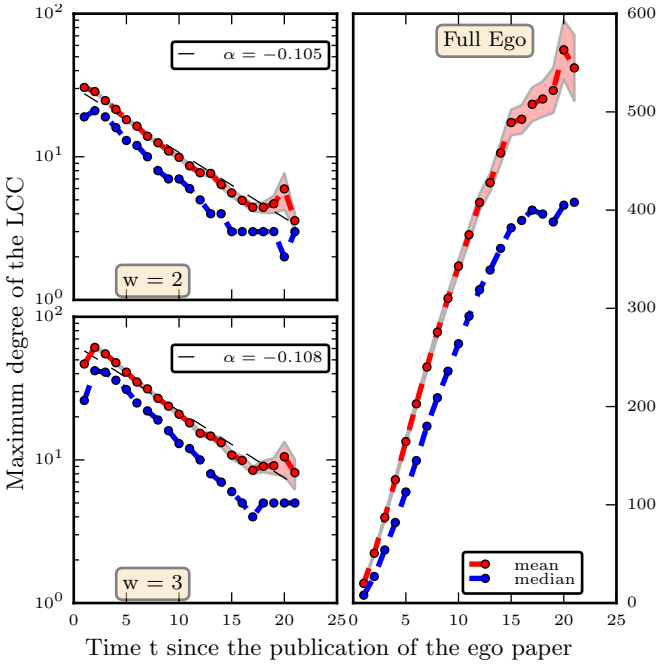

Figure A3: Time Evolution of mean and median relative size of the maximum degree of the lcc for $W = 2$ (top) and $W = 3$ (bottom) (left) along with an exponential fit and of the full EN (right).

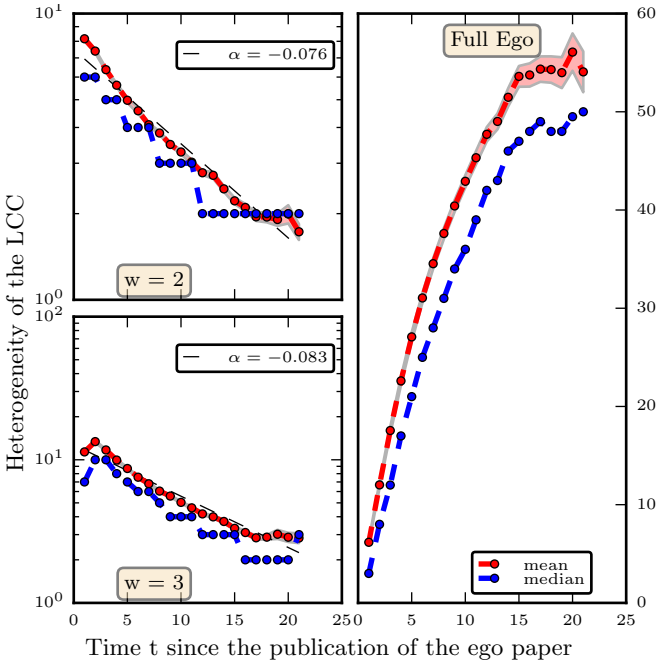

Figure A4: Time Evolution of mean and median relative size of the degree heterogeneity of the lcc, definied as $\frac{\sum d^2}{\sum d}$ for $W = 2$ (top) and $W = 3$ (bottom) (left) along with an exponential fit and of the full EN (right).

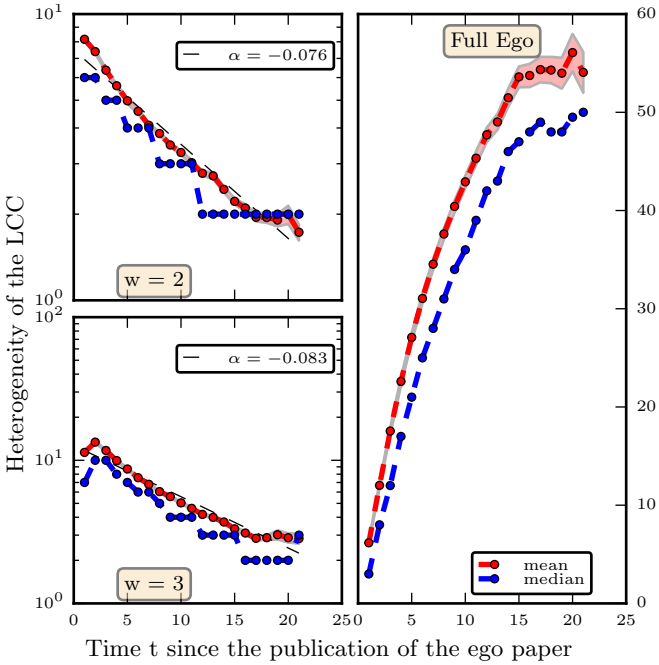

Figure A5: Time Evolution of mean and median relative size of the absolute size of the lcc for different citation volumes: $500 \leq c \leq 1000$ (left), $1000 \leq c \leq 2000$ (center) and $c > 2000$ (right). he higher the citation volume, the faster the decay. This seems to be caused by the fact that the time required for the network to collapse is identical and thus curves starting from higher values (linked to higher citation volumes) inevitably need to fall faster.

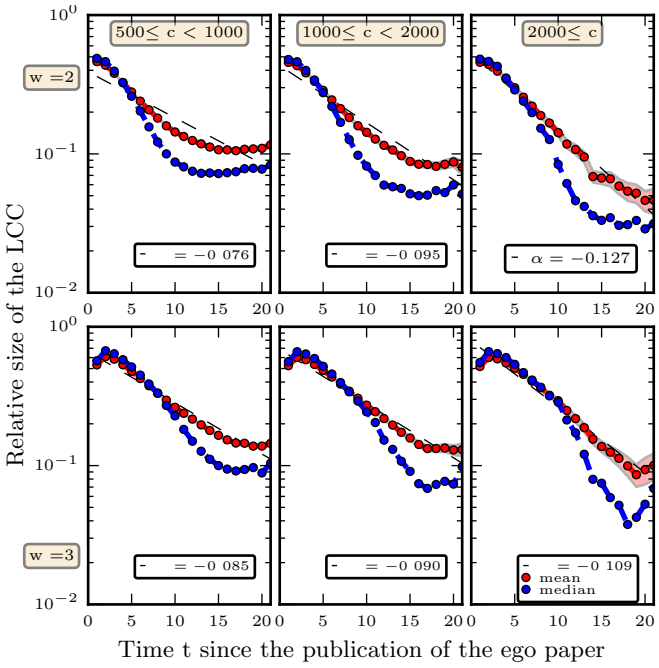

Figure A6: Time Evolution of mean and median relative size of the absolute size of the fraction of references that stay within the EN for different citation volumes: $500 \leq c \leq 1000$ (left), $1000 \leq c \leq 2000$ (center) and $c > 2000$ (right).

Publication 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.

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

Pietro della Briotta Parolo,[1] Kimmo Kaski,[1] and Mikko Kivelä[1]

[1] *Department of Computer Science,Complex Systems Unit, Aalto University School of Science,Finland*

(Dated: June 8, 2017)

The dominant paradigm in tracking the flow of scientific knowledge is to count direct citations between published articles. However, scientific articles are built on articles they cite which in turn are based on cited articles. Similarly, the knowledge created in an article is not only retained by articles that directly cite it, but it persists through chains of citations. Here we investigate this cumulative knowledge creation process by using two stylized models of knowledge flow and a citation network of around 35 million publications. We show that the persistent influence of papers in the global scientific corpus is positively correlated with the citation counts but that there is a large variation in the influence values of papers with similar citation counts and publication dates. It turns out that the papers related to Nobel Prizes are over-achievers in terms of persistent influence when compared to papers with similar numbers of direct citations. We also identify articles with very high persistent influence as compared to the citation count and find that many of such articles are early works that eventually lead into development of hot research topics of their time. Finally, we investigate the diffusion of knowledge across various scientific fields and between them we find large variation in the rates at which they share knowledge with each other. Note that these rates have been systematically increasing for several decades. However, we observe that this trend the rate at which publication volumes increase.

I. INTRODUCTION

Since the seminal work of de Solla Price [1] quantitative analysis of knowledge spreading through scientific publications has become a matter of great interest. The analysis of bibliometric data not only allows to shed light on the structure of science and its knowledge accumulation, but also subsequently get insight into citation distributions [2–4], collaboration networks [5, 6], geographical patterns of collaborations and citations [7–9], as well as to grasp the structural changes that take place at the level of scientific fields [10–12]. In this line of research the citations between scientific publications are of paramount interest. Citations can encode various meanings between articles and be based on different conceptual foundations [13], but perhaps most often they indicate that some knowledge from the cited article is being used in the citing article [14]. The citations can be considered as networks that can be used to investigate the evolution of science and the spreading of knowledge in a wider scope than just between individual publications.

Until now the main research paradigm in science of science has been to focus locally on the direct citations between two articles. This thinking is exemplified by the quality measures that are based on direct citations, such as the H-index[15], the Journal Impact Factor [16], and many others. Even though it has been pointed out that these methods have structural limitations [17–19], the standard response has been the one to circumvent these limitations by developing specific adjustments [20–23]. However, virtually all of these measures are still based on local analysis, i.e. the count of direct citations received by the publication/author/journal whose rank one is trying to determine. This local paradigm is in contrast with the structure of science itself, because science is a cumulative process where researchers are always "on the shoulders of giants" and in which one's results are intrinsically based on massive amount of previous work, not just the articles that are directly cited [14, 24]. Therefore, when attempting to study the structure and behavior of scientific knowledge accumulation it is necessary to look at the whole process and not just focus on a local area of the system.

The reason behind the local viewpoint is presumably its simplicity, and the existence of and access to local data; In order to track flows further away one needs to have large-scale global data on citations between publications. Some previous work has been done to looking at the impact of citations beyond the local perspective, e.g. PageRank-type algorithms with diffusion combined with teleportation have been used to rank publications [25] and individual scientists [26], where some interesting outliers have been found in these studies even though the local and global measures correlate positively. One study instead looked into the in-component structure of individual papers in citation networks, and showed how the link to under-cited work of Nobel-Prize winners have high ranks within the global network of articles [27]. Others have attempted to envision the network of scientific papers as a platform on which an idea can spread [28] or as system similar to social media in which ideas replicate from one publication to the other [29]. In this work we start from a similar perspective of global-scale analysis of citation networks, but with the idea of modelling the flow of knowledge within a large citation network. In particular we want to answer the question: starting from a paper or group of papers, where and how does the information or learned knowledge flow in a citation network if one looks beyond the direct citations?

In order to answer this question we use a massive inter-

disciplinary dataset of articles citing each other and unlike previous studies this comprehensive citation network allows us to extend the analysis to interactions between most fields, subfields and thousands of scientific journals. We focus on studying the spreading of knowledge, originating from a certain seed of papers, through the network based on citations. This is accomplished by selecting a starting article, or a starting group of papers coherent in terms of publication year and scientific field. We then spread the knowledge to future articles from the initial papers through the citation network by following chains of citations from cited papers to citing papers. Different from the PageRank-type of algorithms we are not focusing on anonymous flows of information, but we track the starting point of each unit of knowledge. By doing this we can show how the spreading of scientific knowledge between different fields, subfields, journals, and individual papers takes place and how it changes or evolves in time.

This paper is organised as follows. We first describe the citation data used throughout this study. Then we introduce two stylized models of knowledge spreading. The first one, which we call *persistent influence* is used to track the amount of influence the individual publications have on others, and the second one is used to analyse the *diffusion* of knowledge between fields and other groups of publications. The persistent influence is compared to direct citations of individual papers, and we especially focus on paper associated with Nobel Prize winners and articles that have large discrepancy between direct and indirect influence. For the diffusion process we focus on the speed at which knowledge spreads out of the seed field or subfields, and at how this speed has changed over the years.

II. DATA DESCRIPTION

We use the data set that consists of all publications (articles and reviews) written in English from the year 1898 till the end of 2013 included in the database of the Thomson Reuters (TR) Web of Science. The data set contains a journal assignment for most publications and most journals are further assigned to one or more subfields. We filter out articles and journals for which these information is not available, which leaves us around 35 million publications in around 15 thousand scientific journals. We further map the subfields of the publications into major scientific fields [28].

We use the above filtered set of citations between articles to construct a network where there is a link from citing article to the cited article. We use the publication time information of the articles to remove links where the date of the cited article is not earlier than the date of the citing article. In total we remove 1.7% of the links this way, and we are left with a citation network without any cycles (*i.e.*, a directed acyclic graph). To avoid boundary effects for the latest articles, for which most articles citing them are not in our data set, we only consider the nodes in the citation network until the year 2008. Previous literature [30] shows that the typical life cycle of a publication in terms of citation is completed within 5 years from date of publication. Because the data used here ends in 2013, limiting our attention to articles published until 2008 minimizes the boundary effects originating from missing data on future articles.

III. PERSISTENT INFLUENCE PROCESS

We next want to track how the knowledge created in an article percolates through the network of articles. It is difficult to measure or quantify the amount of knowledge in scientific articles and their origin, so we have to do some simplifying assumptions. First, we assume that each publication is only using information that is present in the articles it cites. Second, in absence of better information, we need to assume that each of the cited articles are equally important for the citing article.

We can formalize the above ideas in a simple persistent influence spreading process. Starting from an original seed publication s we attribute to it an initial value of influence $I_s = 1$, while all other publications have an initial value of 0. We then update the influence values of article published after the original one in chronological order such that node j pulls scientific influence, if present, from all articles it cites and updates the persistent influence that the seed article has on it:

$$I_j = \sum_{i \in N_j} \frac{I_i}{k_j^{in}} \tag{1}$$

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 normalisation guarantees that the sum of influence that the cited articles have on article j is constant and that the influence value never exceeds 1. When the process continues, the influence values dilute but at the same time it is spread to increasing number of articles.

In this pulling mechanism we consider the relative influence of the cited paper on the citing paper such that influence of the cited article is passed on to the citing one such that each citation in the reference list has the same importance. A hypothetical publication with only one reference will draw all its influence from the cited paper as its scientific results are entirely based on that previous work in our model. Similarly, an article that is cited by a review article which also cites hundreds of other publications has to share the attention with all of the other references, and only a small fraction of the information present in the cited article is influencing the review.

The persistent influence values I_i can also be considered in terms of a diffusion process that goes backwards in time. Consider a random walk where starting from an article i one at each step selects another paper uniformly randomly from its reference list and jumps into that article. This procedure is then repeated until we reach an

article that has an empty reference list. The probability that this random walker visits article s when starting from i is exactly the persistent influence I_i that article s has on i.

One can gain further intuition on how the influence spreading process works by considering how papers influence scores would develop in a simple citation model. Consider a citation network where articles are published in generations and they only cite articles in the previous generation[31]. Further, the number of articles in each generation $n_t \to \infty$ such that the rate of change $\mu = n_{t+1}/n_t$ is constant, and the in- and out-degree distributions have only degrees that are small compared to the system size n_t but are otherwise arbitrary. If the number of citations a paper receives and gives out (out-degree and in-degree) are independent then the sum of influence of all papers in generations t follows on average $I^{(t)} = \mu p(k^{in} = 0) I^{(t-1)}$, where $p(k^{in} = 0)$ is the probability that a node has zero in-degree (i.e., it receives no citations). That is, the total influence of a paper to all future research remains constant in the case where the systems size is also constant and there are no "dead-end" articles. However, in reality the scientific input has been continuously growing [28], and we expect that on average the influence of early papers will be growing. This is of course only the average picture and some papers influence will be dying out and others growing faster than average.

A. Results

We will now apply the persistent influence process to the data set of publications and their citations described earlier, and use this simplified process to gain insights on how publications have had impact on scientific world at different stages. We start by a detailed description of a single sources persistent influence as an example, and then continue by looking at large sets of source articles.

Fig. 1 shows the persistent influence profile of Roy J. Glauber's seminal paper on photon correlations [32] published in 1963, which eventually led to him winning a Nobel Prize. We compare results from the full global persistent influence process described above to a more conventional local analysis where the influence or impact of a paper is determined only by the direct citations. In order to make these two comparable we use the Eq. 1 also for the local influence scores, but disregarding everything else but direct citations to the source article. The main difference between the global and local profiles at first sight is that the global profile is much smoother than the local one.

The influence of the source paper on individual fields displayed in Fig.1(a) shows a strong persistence in Physics (dark red) with a gradually growing contribution from two other fields (dark and light orange), which is already getting stable after 10 years. This is effect is not visible in the local profile even though the same fields start to cite the seed article much later on. This phenomenon is either due to the propagation in the network of the initial "sparks" received in the immediate time after publication (cfr. panel b) or due to the following layers of citations. Looking at the global development of subfields (panels c-d) we can see a similar behaviour, which is represented by the steady growing of optics (dark red). In the local picture only a few citations from such subfield target the paper, but in the global persistent picture the process leads to optics becoming the most relevant subfield, an event which will take place in the local pictures only after 20 years from publication. Similarly, the contribution to Optics is mainly due to two journals: *Physical Review A* and *Optics Communications*, which once again are not present in the local profile. The total persistent influence of Glauber's article on articles published during each year grows (with the exception of few years), but remains relatively stationary when only direct citations are considered.

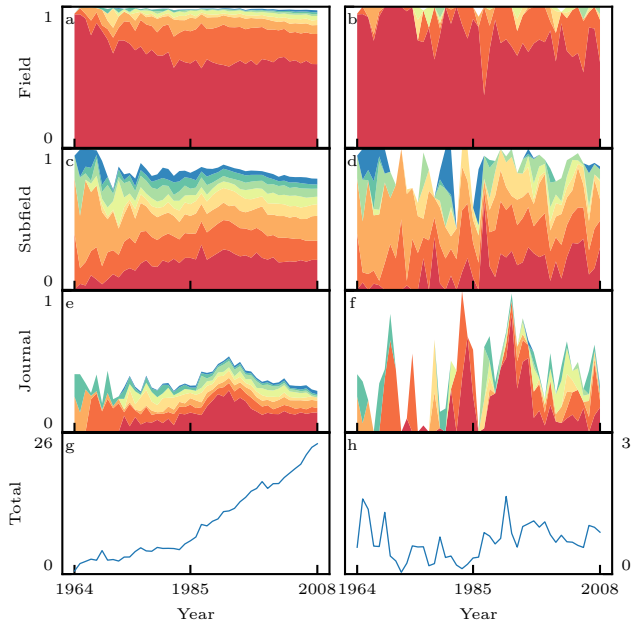

FIG. 1. Global and local influence profiles of Glauber's Nobel Prize winning paper from 1963 to 2008. The left column (panels a, c, e, g) shows the influence measures linked to the global process, while the right column (panels b,d, f, h) shows the influence calculated only from direct citations. Panels a and b show the relative distribution of influence among fields with most influenced fields Physics,Mechanical and Engineering (from top to bottom in this order). Similarly, panels c and d (e and f) show the influence on subfields (journals) with the most influenced subffields being Optics, Physics, Multidisciplinary Physics (Phys. Rev. A, Phys. Lett. A, Phys. Rev. Lett.). Only the contribution of 8 largest fields/subfields/journals are shown and the rest is shown as white space. The bottom panels g and h show the total influence, i.e., the sum of influence values I_i for articles published in each year.

We calculated the influence values for all papers in our

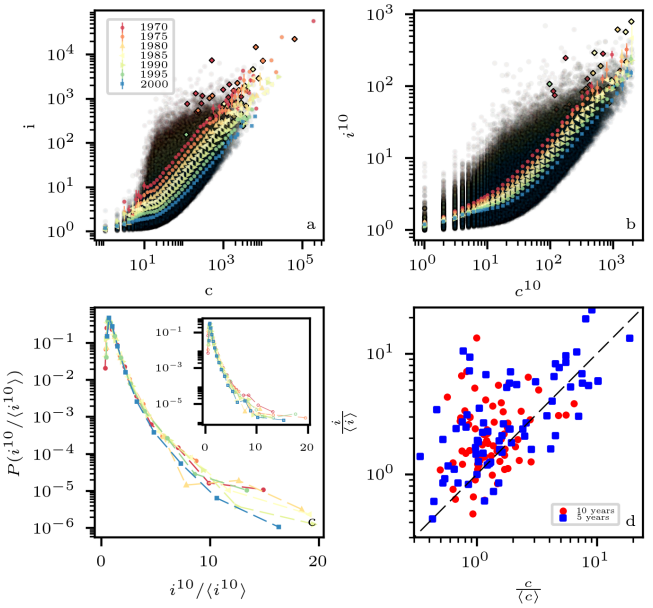

FIG. 2. Panel a shows the scatter plot between citation values and influence values for different years along with bin averages (filled markers). The diamond shapes represent Nobel papers, while the color is chosen according to the rounded closest value among the available years. Panel b shows the same information, but with data calculated after 10 years. Panel c shows the distribution of influence values divided by the average for a group of papers with similar order of magnitude of citations (≈ 100). The subpanel shows the same distribution for one order of magnitude less. This corresponds to the distribution of influence values for vertical slices of panel b. Panel d shows a scatter plot for the outperformance of Nobel papers both in terms of total citations and influence, compared to similar paper in terms of number of citations after 5 (blue squares) and 10 years (red dots). Dots on the right side of the diagonal indicate Nobel publications whose citation performance is higher than the influence one. While there are cases in which the Nobel papers are underperforming, they're roughly twice more likely to underperform (i.e. to have values less than the average) citation wise (30%) rather than influence wise (13%).

datasets with at lest 20 citations, published between 1970 and 2008. In total this amounts to around $6,2$ million publications. Out of this set of article we selected 74 papers that were associated with a Nobel Prize [33]. We focus on the total influence values on each year and the cumulative influence values where the sum of all influence values are summed up to a given number of years after the seed publication. The results are summarized in Fig.2.

There is a positive correlation between total number of citations and total influence an article receives (Fig.2a), and this relationship resembles the results in [25], with a strong correlation between the two values, but at the same time showing a huge variance of influence within articles that receive similar numbers of citations. Especially in the central range of citations (10 to 100) the distribution of influence can span numerous orders of magnitude. This indicates, as expected, that the number of citations per se is not sufficient to summarize fully the influence that a single paper has had in the scientific community. Also, we see a clear advantage of older papers, which manage to gather a significantly higher amount of impact with the same number of citations. This is to be expected as more recent papers have had less time to gather impact among their scientific offspring and thus suffer of a lag in their impact pattern. Interestingly, we see that papers associate with Nobel Prize fall in the top right corner of the figure, indicating high values of both citations and influence. However, there are clearly a group of Nobel publications that are well above the average when compared to other papers with similar publication year and citation count. This is coherent with the fact that one of the main reasons behind a Nobel Prize winning discovery is a significant contribution to the scientific world.

The older paper have had more time to gather total influence in our data set, and to remove this advantage we calculated the cumulative influence value and citation count after 10 years from the publication date of each article (Fig.2b). A similar relationship as described earlier between the influence and citation count exists also in this case. Even the variance between influence values of articles with similar numbers of citations remains high (Fig.2c and insert). This shows that also on a shorter time scale, publications with very similar number of citations, manage to havedifferen a extremely varied impact in the scientific world. Once again, papers related to Nobel Prizes show to be overachievers, having influence also in this time scale significantly higher than the average for their number of citations. The difference shown in the average value between papers in different publications year is less strong than before and could be caused by an increase in the length of reference lists, which cause the denominator in Eq.1 to reduce the amount of impact in the citing papers. In fact, when comparing the influence distributions of articles with similar number of citations across years (Fig.2c) we can see that, across decades, the relative distribution is super-exponential and relatively stable, indicating that despite a change in average total influence values after 10 years (i^{10}), the distribution relative to average remains constant in time. Fig. 2 (d) shows the correlation of the outperformance of Nobel papers in both influence and citations. For each Nobel paper we return the total influence and citation values of papers within 10% of their citation count after 10 or 5 years and published in the same year. We then proceed to calculate the outperformance value of the Nobel paper by calculating the ratio between its own total influence and citation value and the average of each distribution for papers in the same citation bin. The figure shows the correlation between the outperformance in impact and in citations. In 73% of the cases (54 papers out of 74) the citation outperformance has been greater than the impact one. Also, the average impact outperformance

after 5 years is 41% greater (4 versus 2.8), while after 10 years it rises to 66% greater (2.44 versus 1.47). Finally, while there are cases in which the outperformance has been greater in terms of citations, there are no extreme values in bottom right corner (high citation performance, low influence performance), while there is a solid vertical band in the opposite side of the plot, indicating a group of papers with low citation performance and high influence instead. This shows that Nobel papers, on average, outpeform papers who had similar citation counts after the same amount of time. This is due to the continued ability of the original Nobel paper (as well of its scientific offspring) to continuously propagate their ideas in the scientific community, as one would expect from such important publications. This result supports the finding of [27], where groundbreaking papers by important Nobel laureates were found to have a more "compact" network of child nodes over multiple layers.

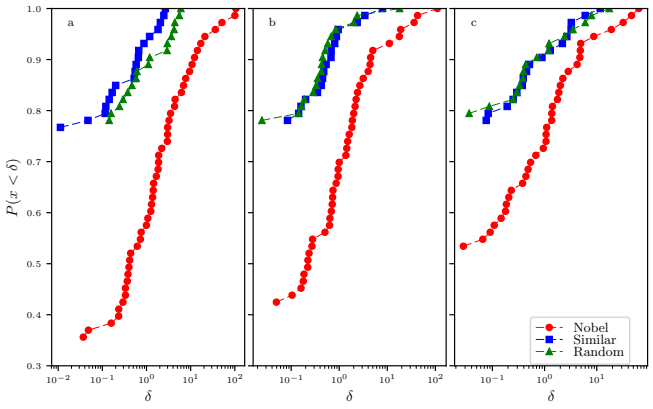

FIG. 3. Cumulative distribution of relative difference in persistent influence rank and citation count rank δ (given in Eq. 2). Three categories of papers are considered: Nobel papers (red circles), Nobel control set with papers within a 3% in citation volume in the same time interval (blue squares), and for random papers (green triangles). In case no papers were found within the 3% interval, the most similar paper in terms of citation was returned. The different panels show the values of δ for rankings calculated in different times, with panel a being the ranking in 2008, panel b being the ranking after 10 years and c the ranking after 5 years.

In order to better study the difference in performance between citation counts and persistent influence, we defined a measure that quantifies the relative difference in rankings for each paper in each category. In order to do so we took all publications within the same year and proceeded to rank them both according to citation count and persistent influence. We define the relative change in ranking is as

$$\delta = \frac{R_c - R_I}{R_I}\,, \qquad (2)$$

where R_c and R_i are ranks terms of citations and persistent influence such that small rank means high value. A negative δ indicates lower influence ranking than citation ranking. The division by *iRank* guarantees that papers with high influence articles receive a higher δ compared to a paper that has the same change in rank but small influence value. This definition allows us to compare outperformance levels across years, removing the bias of higher counts for older papers. Fig.3 shows the cumulative distribution of δ at different times for the same 74 Nobel papers, a random selection of papers within 3% of citations of each Nobel paper and randomly selected papers. Nobel papers have, on average δ 50% bigger than similar papers (6 vs 4), showing that they are more likely to have high influence as compared to citation count. In general, we can see that the performance of these similarly cited papers is more similar to the one of randomly selected papers than the one of the Nobel publications. Furthermore, it appears that while the number of papers with positive δ is constant for the control papers, it is significantly increasing for the Nobel publications, indicating that Nobel papers are more likely to climb the influence rankings as the time window increases and as their cumulative contribution to science expands.

Table I shows articles with largest change in δ for all papers in our dataset. That is, these are papers that have received relatively few citations but have high influence scores, which means that they might have contributed to the development of science by inspiring further research even though they are not well cited. It is interesting to note how many of the papers in the first positions in the Table I are from the 70s and are linked to the field of Genetics. The first one has been cited by EM Southern's work on the *Southern Blot*, a method used in molecular biology for detection of a specific DNA sequence in DNA samples, while the second, third and fifth are in the reference list of Sanger's Nobel Paper *DNA sequencing with chain terminating inhibitors*. This gives also information about the massive impact that the sequencing of DNA has had on the whole scientific world. However, the most striking feature is the diverse aspects of science that this list includes. #5,#9,#15 and #21 are all linked to the identification, classification, or prediction of very well known diseases (Prostate Cancer, AIDS, Leukemia). We can also find many papers from physics, with #4,#12 and #24 being linked to the discovery of High Temperature Superconductivity, while #11 and #30 are linked to the development of Carbon Nanotubes and, in general, of Material Science. #10 is among Amano's works that lead to his Nobel Prize for the invention of efficient blue light emitting diodes. #25 is a small summary of the recent (at the time) discoveries in the mathematical field of Fractals, which was among the few cited works in the famous *Self Organized Criticality: An Explanation of 1/f Noise*. Also, we can see contributions from Economics and Engineering with #27, which discusses a computer method able to improve the efficiency of production of industrially assembled products. #28 is one of the earliest attempts of statistical methods for assessing agreement between different clinical measurements.

R_δ	R_c	R_I	Year	Title
1	37588.5	5	1974	Hybridization On Filters With Competitor Dna In Liquid-Phase In A Standardand A Micro-Assay
2	23366	5	1976	Nucleotide And Amino-Acid Sequences Of Gene-G Of Phix174
3	62269	18	1975	Invitro Polyoma Dna-Synthesis - Inhibition By 1-Beta-D-Arabinofuranosyl Ctp
4	88381	32	1980	Inhomogeneous Superconducting Transitions In Granular A1
5	26353.5	10	1997	An Adjustment To The 1997 Estimate For New Prostate Cancer Cases
6	28047	11	1970	Molecular Hybridization Between Rat Liver Deoxyribonucleic Acid And Complementary Ribonucleic Acid
7	63260	25	1985	A Novel Method For The Detection Of Polymorphic Restriction Sites By Cleavage Of Oligonucleotide Probes - Application To Sickle-Cell-Anemia
8	105590.5	42	1983	Phase-Diagram Of The (Laalo3)1-X (Srtio3)X Solid-Solution System, For X-Less-Than-Or-Equal-To 0.8
9	131750	72	2004	A New Method Of Predicting Us And State-Level Cancer Mortality Counts For The Current Calendar Year
10	114723	67	1988	Zn Related Electroluminescent Properties In Movpe Grown Gan
11	26020	16	1989	Structure And Intercalation Of Thin Benzene Derived Carbon-Fibers
12	12231.5	8	1985	The Oxygen Defect Perovskite Bala4Cu5O13.4, A Metallic Conductor
13	19801	13	1972	Translation Of Encephalomyocarditis Viral-Rna In Oocytes Of Xenopus-Laevis
14	4216.5	3	1974	Amplified Ribosomal Dna From Xenopus-Laevis Has Heterogeneous Spacer Lengths
15	42143.5	30	1975	Classification Of Acute Leukemias
16	62485.5	48	2002	Wild Topology, Hyperbolic Geometry And Fusion Algebra Of High Energy Particle Physics
17	58242.5	46	1978	Relation Between Mobility Edge Problem And An Isotropic Xy Model
18	42981.5	34	1986	Transcriptional And Posttranscriptional Roles Of Glucocorticoid In The Expression Of The Rat 25,000 Molecular-Weight Casein Gene
19	114240	91	1986	The Use Of Biotinylated Dna Probes For Detecting Single Copy Human Restriction-Fragment-Length-Polymorphisms Separated By Electrophoresis
20	92031	74	1989	A Solid-State Nmr-Study On Crystalline Forms Of Nylon-6
21	89271	74	1982	Multiple Opportunistic Infection In A Male-Homosexual In France
22	11535	10	1970	Synthesis Of Ribosomal Rna In Different Organisms - Structure And Evolution Of Rrna Precursor
23	11227.5	10	1999	Small-World Networks: Evidence For A Crossover Picture
24	12064.5	13	1987	Superconductivity At 52.5-K In The Lanthanum-Barium-Copper-Oxide System
25	71919	82	1986	Fractals - Wheres The Physics
26	28044	33	1986	The Complete Structure Of The Rat Thyroglobulin Gene
27	21222	25	1979	Interference Detection Among Solids And Surfaces
28	81374	99	1979	Comparison Of The New Miniature Wright Peak Flow Meter With The Standard Wright Peak Flow Meter
29	7962.5	10	1972	Studies On Polynucleotides .105. Total Synthesis Of Structural Gene For Analanine Transfer Ribonucleic-Acid From Yeast - Chemical Synthesis Of An Icosadeoxyribonucleotide Corresponding To Nucleotide Sequence 31 To 50
30	26908.5	34	1992	Materials Science - Strength In Disunity

TABLE I. Publications with highest relative difference in influence rank and citation count rank δ (given in Eq. 2).

The list also shows evidence of the relatively new field of complex networks, with #23 being among the earliest papers in the field, being cited by virtually all the most significant later publications in the field. Globally, we can see how δ is able to grasp the growth in the whole scientific field of certain discoveries/subfields/hot topics by being able to identify low cited papers that have been crucial in their early stages.

IV. DIFFUSION

The persistent influence spreading method we just introduced is a simple and elegant method to track the spreading of knowledge in the citation network, but it is not the only plausible one. In the influence spreading the tracked quantity can be copied and the total amount in all articles can grow. Next we instead define a diffusion method where the original mass placed on the seed node (or nodes) is always strictly conserved. This allows us to track the diffusion of ideas not only from single articles but accross journals, subfields, and fields.

The idea behind the diffusion method is to start a random walker from a seed article that is randomly selected from a set of seed articles, and then, at each step, let it move from an article to future article citing it. Note that this process would be sensitive to the time window we choose, as future articles that we do not know about yet would change the probabilities of trajectories of the walkers as they introduce additional citations to older papers. To negate this effect, we will focus on walks that have not passed beyond our observation year. That is, we only use the information available in the observation year and the random walk process is recalculted for each starting year and observation year pair.

We initialized a set of N seed papers to which we have assigned the same initial value $v_i = 1/N$. We also tested assigning initial values asymmetrically by looking at how many citations each paper has received in the first 5 years (plus one, to take care of citationless papers): $v_i = \frac{1+c_i^5}{\sum_j (1+c_j^5)}$. This count acts as a proxy of how successful the paper has been in general, but this alternative initialization strategy resulted in qualitatively similar results to the more simple strategy and we do not show these results here.

Similar to the initialization of the diffusion process, we tried out two different ways of selecting the probabilities that the random walker uses to follow citations to the future. In the simple case the walker jumps from the cited article to each citing article with the same probability, and in the other case the random walker preferentially chooses articles that receive more citations in the coming five years c_i^5. The results for both processes are similar and here we only show them for the simpler process. For technical details on how the process was made computationally tractable and implemented see Appendix A.

A. Results

When starting the diffusion process, we select a field, subfield, or a journal and a starting year, and track to which fields, subfields, and journals the probability mass for the diffusion process ends up in each year. Fig 4a-c shows examples of typical results. Here we have chose the initial field of Economics, the subfield of Evolutionary Biology and the British Medical Journal. As we can see, the initial values starts from a high value, which is not exactly one, as we include also papers published in multiple fields (in that case we split the value equally among all fields or subfields the paper has). As time goes by, the "scientific value" diffuses out of the initial group and we can see that other minor groups get increasingly more relevant, with the initial field still losing value, but at a slower pace. It seems also that the difference in pushing methods do not play a role, whereas the initialization method based on citation does, making scientific value of the original field fall faster.

In order to study the change of value within the ini-

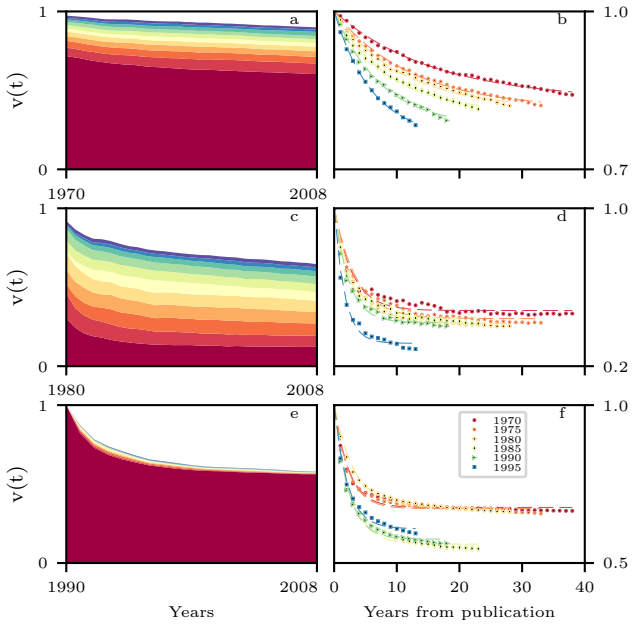

FIG. 4. Example of the diffusion method (a-c-e) and data fitting (b-d-f). The diffusion of scientific value for (a, b) Economics in 1970, (c, d) Evolutionary Biology in 1980, and (e, f) the British Medical Journal in 1990. The darkest tone of red shows always the amount of scientific value retained by the same initialization group (field, subfield or journal), while the other colors show the 10 next fields/subfields/journals with the highest combined scientific value across time. For panel a the colors (from bottom to top) represent the following fields: Economics,Management, History, Political, Mathematics, Geosciences, Social Sciences, Agriculture, Environmental, Sociology, Multidisciplinary. For panel c the colors represent the subfields: Evolutionary Biology, Biology, Miscellaneous, Plant Sciences, Anthropology, Ecology, Zoology, Genetics & Heredity, Arts & Humanities, General, Biology, Entomology, Biochemistry & Molecular Biology. In panel e the colors represent the journals: Br. Med. J., Lancet, Br. J. Gen. Pract., Bmj-British Medical Journal, Med. J. Aust., Postgrad. Med. J., Arch. Dis. Child., J. Clin. Pathol., Med. Clin., Soc. Sci. Med., Ann. R. Coll. Surg. Engl. The right column panels instead show the renormalized value of v retained within each field/subfield/journal for different years and with the exponential fitting (dashed line).

tial fields, subfields and journals we have looked only at the amount of value retained by each group. For each start year we take the yearly values and we renormalize them with the initial value of the field, so that the history shows the relative amount value retained. We then proceeded to fit each curve with an exponential of the form: $v(t) = (1-\beta)e^{-\alpha t} + \beta$, which follows well the typical shape we observe for the curve (see Fig. Fig 4d-f). This allows us to quantify numerically both the rate of change of the value in the initial years (through α) and find the final plateau value (through β). Therefore, α can be used to measure the speed at which one field shares its knowledge with other fields, and β instead represents the intrinsic "conservativeness" of a field, i.e. the amount of knowledge retained within the boundaries of the field itself in medium time scales. In order to provide an easier metric for the decay we can introduce a related parameter called half life $t^{1/2}$ defined as the time required to lose half of the possible plateau value $1 - \frac{1-\beta}{2}$:

$$1 - \frac{1-\beta}{2} = (1-\beta)e^{-\alpha t^{1/2}} + \beta\,, \tag{3}$$

which allows us to use the conventional definition of half life:

$$t^{1/2} = \ln(2)/\alpha\,. \tag{4}$$

Overall, we can see that the same general pattern is retained: the probability mass that remains within the same initial scientific area, may it be field, subfield or journal has a sharp initial decay, followed by a plateau. Also, subfields and journals seem to have the same property as fields, i.e. having a faster diffusion of scientific ideas to other "competitors" at a faster rate in time.

With these ideas in mind we can try to put together the information about all possible fields, subfields, and journals. Table II shows the values for the half lives and βs in 1970 and 1990 for equal initialization and pushing. We can see that in general there is a decreasing trend for half lives while the plateau value β instead shows a much more stable pattern. It is also interesting to point out some patterns for individual fields. We can see that the field of multidisciplinary has the lowest half life for both starting years, coherently with the fact that it is meant to be a field open to sharing its knowledge with others. However, the change in β is positive and the second highest (behind Music), indicating that nowadays the field tends to retain more value to itself, coherently with the evidence that shows the increasing role of interdisciplinarity in science [10, 11, 34, 35]. It is also interesting to note that while in 1970 we can see some humanistic fields showing very high values for their half lives (Phylosophy, History, Anthropology, Literature, and Linguistics), these fields also show some of the highest changes in time, putting them much closer to hard sciences in modern days than they were before.

Fig. 5(a-c-e) shows the change of half life for some of the fields, all subfields, and a list of journals. We can see in panel a that all fields show a speeding-up pattern, losing between 20 and 60 percent of their half-life values, while for subfields and journals (panels c and e) the more recent cumulative distributions of half-lives are above the older ones, showing that the values have decreased on average.

Previous studies show [28] that measuring the time in years might not be the best choice to measure the rate at which changes happen in science, and instead one should use the numbers of papers published as a better metric. The idea is that the system is "updated" (i.e. scientific value is propagated) every time a new publication is

	1970		1995		Δ	
Field	$t^{\frac{1}{2}}$	β	$t^{\frac{1}{2}}$	β	$t^{\frac{1}{2}}$	β
Philosophy	19.7	0.84	4.36	0.90	-78%	+3%
Economics	11.0	0.83	4.20	0.76	-62%	-8%
Psychology	8.93	0.72	3.44	0.67	-61%	-7%
Linguistics	8.86	0.87	3.02	0.90	-66%	+3%
Chemistry	8.55	0.80	1.99	0.80	-77%	0%
Music & Dance	7.83	0.82	6.18	0.98	-21%	+2%
Gen. Humanities	7.25	0.85	3.43	0.95	-53%	+12%
Mathematics	7.14	0.87	3.21	0.79	-55%	-9%
Medicine	6.54	0.83	3.20	0.85	-51%	+2%
Sociology	6.34	0.80	3.72	0.73	-41%	-9%
Engineering	4.89	0.82	2.33	0.79	-52%	-4%
Law	4.38	0.92	7.21	0.80	+65%	-13%
Social Sciences	4.38	0.73	2.35	0.59	-46%	-19%
Physics	4.01	0.82	2.32	0.81	-42%	-1%
Management	3.72	0.78	3.60	0.66	-3%	-15%
Biology	3.43	0.71	1.69	0.70	-51%	-1%
Multidisciplinary	1.33	0.59	1.08	0.59	-19%	18%

TABLE II. Half-lives in years ($t^{\frac{1}{2}}$) and asymptotic fractions (β) for a subset of fields in 1970 and 1995 and for equal initialization when the evolution of the diffusion process is fitted to Eq. 3 along with the relative change for each value.

introduced in the system. Furthermore, while the number of publications grow exponentially, the growth rate is sufficiently small: $N(t) \approx N_0 e^{\delta t}$ with $\delta \sim 0.05$ across all fields, allowing us to keep the same functional forms for the exponential fits we did earlier:

$$\begin{aligned} v(N) &= (1-\beta)e^{-\alpha N_0 \int_{t_0}^{t_N} \exp^{\delta t}\, dt} + \beta \\ &\approx (1-\beta)e^{-\alpha N_0 \left[\frac{1+\delta t}{\delta}\right]_{T_0}^{T_N}} + \beta \\ &\approx (1-\beta)e^{-\alpha \delta^{-1}(N_0(1+\delta(T_N-T_0))} + \beta \\ &\approx (1-\beta)Ke^{\alpha N_0 \Delta(T)} + \beta \\ &= (1-\beta)Ke^{\alpha^* \Delta(T)} + \beta\,. \end{aligned} \tag{5}$$

Therefore, we are able to quantify the half life not in terms of years, but rather in terms of number of published papers. For simplicity, we decided to use for each field the number of papers published in the field as a renormalizing measure, while for subfields and journals we used the data from all scientific publications.

Fig.5(b-d-f) shows the half lives which are renormalized such that they are given in terms of published papers, as described in Eq.5. Interestingly, the decreasing behaviour is no more dominating, with only one field (Chemistry) showing a downward pattern. All other largest fields, instead, either remain constant or show a significant increase in their half lives over time. The same can be seen in the distribution for subfields and journals, with the previous color order being now inverted, indicating that the renormalized values are increasing in time on average.

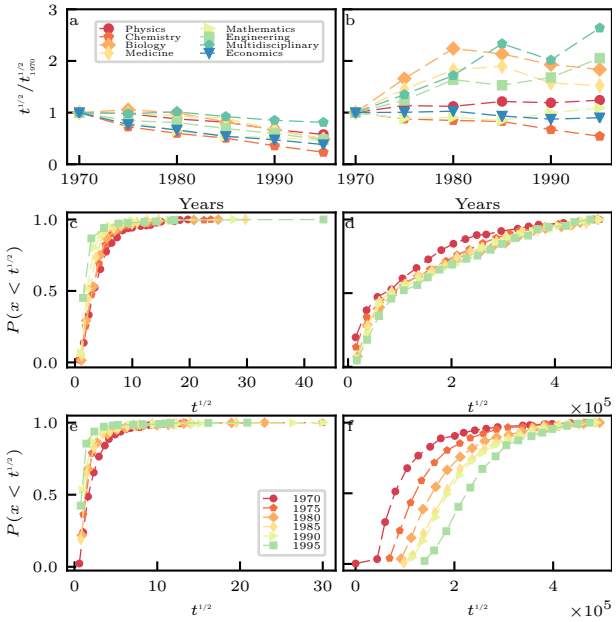

FIG. 5. Changes in half life in time for the regular (left column, panels a-c-e) and renormalized scenario (right, panels b-d-f) and for different grouping of papers. Panels a and b show the evolution of the half life for 8 different fields in units of the half life they had in 1970 in order to be able to compare the trend. In the regular scenario all fields show a downwards trend, while in the renormalized scneario certain fields show an increase in half life, indicating a slowing down in the time required to share knowledge with other fields, while others remain constant. Panel c shows the cumulative distribution for half lives for all subfields in the regular scenario, while panel d shows the same distribution for the renormalized case. Panels e and f show the same for journals.

V. DISCUSSION AND CONCLUSIONS

Ever since bibliometric data of scientific publications has been available there have been efforts to analyze such data with the goal of quantifying scientic research. The dominant framework has been to use the counts of direct citations between articles, journal, and research fields in order to quantify the relationships between them, and to rank authors [15], publications [36], universities [37], and institutions [38]. The citation counts as measures of quality work reasonably well [39], and apart from few exceptions [40–42], the improvements on these methods have been correcting technical flaws [40–42] instead of focusing on the intrinsic conceptual flaw: these methods work only a limited snapshot of the system, focusing of what is only the first of the many other layers that continuously build on top of each other in the network of scientific publications.

We have introduced two simple methods to analyse how the knowledge created in an article, or in a group of articles, might percolate trough the scientific literature. These methods follow the tradition of modelling

dynamics on networks, where a real observed network is used as a substrate where a stylized models progress is tracked. For example, this approach has been extensively used in network science to study epidemic spreading [43] and social dynamics [44]. In all of these cases the models are not expected to exactly mimic the real behavior, but the goal is to produce behavior that is accurate in the large scale. Our goal in this work was to introduce this approach to knowledge spreading in citation networks.

The first of our two measures, *persistent influence*, is based on citing papers inheriting the knowledge of papers they cite, i.e. the "shoulders" on which they stand on, and therefore propagating the influence of the cited papers, which in turn will be inherited by later papers. As expected, the out-degree, i.e., the direct citation count, is positively correlated with persistent influence, but we also observed that papers that are similar in publication date and citation count have a wide range of influence values. This indicates that the local measure of citation counts can often be a poor proxy for tracking the global cumulative influence of an article. We wanted to test the hypothesis that the discrepancies in the local influence values and global ones are not meaningful but simply noise added by the global process, and to do this we use papers related to Nobel Prizes as a manually curated corpus for high influence on science. We found that the Nobel papers systematically overperform papers with similar citation counts and publication dates in terms of persistent influence, and that the hypothesis that differences in indirect and direct influence are not meaningful is clearly false. Furthermore, we looked at the papers that have the greatest increase in rank while switching from the local to the global scenario, and found that these papers are often early publications in fields that would later become hot topics of their time in the scientific world.

The second knowledge spreading approach we employed was a simple *diffusion* method. We focused on analysing the rate of diffusion of knowledge across fields, subfields, and journals. We found that the curves describing the loss of diffusing knowledge to other fields, subfields, and journals is well described by a common pattern across disciplines. In this pattern the value first exponentially decreases and then reaches a plateau. Each starting time and set of seed articles can thus be described by a plateau value of retained knowledge β and by a typical time required $t^{1/2}$ to share half of the available knowledge. We found that β varies heavily across disciplines, yet remaining constant in time, while the values for the half life, $t^{1/2}$, have been steadily decreasing, suggesting an increase in interdisciplinarity. However, we showed that the increasing speed of information sharing could be explained by the increase in the speed at which publications are produced.

The work done here is forms a basis for future possibilities of the model-based approaches to tracking global knowledge spreading in citation networks. For example, more detailed look on the long term destinations of influence starting from various sources could bring interesting results. Further, one can easily reverse the tracking direction of the persistent influence model and investigate which articles, or groups of articles, in the history have influenced individual papers. With more and more bibliometric data being available, we believe that our findings should encourage future work to analyze science for what it is and has always been: a cumulative process that builds over time in which the successes in scientific discoveries are built on chains of previous successes.

VI. ACKNOWLEDGMENTS

We used data from the Science Citation Index Expanded, Social Science Citation Index and Arts & Humanities Citation Index, prepared by Thomson Reuters, Philadelphia, Pennsylvania, USA, Copyright Thomson Reuters, 2013. This research is partially supported by the European Communitys H2020 Program under the scheme INFRAIA-1-2014-2015: Research Infrastructures, grant agreement 654024 SoBigData: Social Mining. K.K. would also like to acknowledge financial support by the Academy of Finland Research project (COSDYN) No. 276439 and EU HORIZON 2020 FET Open RIA project (IBSEN) No. 662725.

Appendix A: Computational considerations

In order to implement both the diffusion and influence algorithms we had to organize the citation network in Directed Acyclic Graphs (DAGs), as mentioned in the Data Decription section. After this it was necessary to order the nodes in *topological order*. Such ordering guarantees that for every directed edge connecting papers i (the citing paper and j (the cited paper), j comes before i. Such ordering, in principle, should correspond with the time stamp of the publication. However, for older papers we did not have sufficient resolution in the data to rely fully on that, as only the yearly data was provided. Therefore, we took advantage of the facts that papers published in older years are bound to have a lower rank in the order. Thus we built yearly citation networks and ordered them topologically, starting from our latest entry. We then proceeded to arrange the nodes topologically within the year network, building the overall topological order adding one layer of publications at a time. Once a topological ordering of the nodes was created, we built a topological ordering for the edges, sorting them by topological order of the cited paper. This ordering guarantees that each node is visited exactly once and that, while looping through the topologically sorted edgelist, each paper has collected all the value/influence upstream before pushing its own forward. This ordering allowed us to loop through the edgelist only once and to control the start and end of the pushing process by checking for each edge that the topological ordering values of each paper

lie within the year bounds.

In order to implement the diffusion process, we have chosen as seeds the set of all papers being published in the same field in the same year y_{start}. By doing this we are able to select a very coherent set of papers both in terms of subject and time. Once the system has been initialized, the next step is to choose a final year y_{end} as the year in at which we will stop pushing values forward. Hence we loop through the nodelist of all scientific papers in our dataset published between y_{start} and $y_{end}-1$ arranged in topological order in order to initialize the weights for each node. We consider only links to neighbours that point to papers published before y_{end}. Similarly as before, one can choose two methods for initializing the weight of each node i to paper j in its neighbourhood:

- $w_{ij} = \frac{1}{\sum_{k\in N_i} 1}$
- $w_{ij} = \frac{c_j^5}{\sum_{k\in N_i} c_k^5}$

The first definition spreads all the value of each node equally among its child nodes. In the second case instead one takes into account how many citations the child node receives allowing it to get more value the more citations it has received in the next five years.

Once the weights have been initialized we can push the value of each node by looping again through the nodelist in by topological order (this guarantees that no value is ever pushed from a node before the same node has collected all previous value available. The pushing starts from y_{start} and stops in $y_{end}-1$ but spreads to papers published all the way to y_{end}, without pushing any value within y_{end} . This means that we consider as leaf nodes of the system only the first papers to receive value in the final year, as receiving citations in the first year is somewhat hard to obtain (it heavily depends on the month of publication) and one single citation might steal all the value from another paper.

After the pushing has ended we can collect all the values that are left un- pushed in the system. Since the pushing has been carried out by following the topological order of the whole graph, this is simply accomplished by not stor- ing permanently the value of any node that appears in the first column of the edgelist as by definition they will necessarily get rid of all their values. Also, by constructions the sum of the values of all leave sums to one. It is important to notice that in order to collect the data between say 1990 and 2008 one needs to repeat the pushing process for each y_{end} between those years, since the network initialized is different each time. This means that when we collect the data in a certain year, we don't consider what happened in the future (except the 5 year citation proxy). If we were to collect the data in middle years while pushing the values directly to the last year, we would be including links to recent papers that would steal value from the middle years, thus altering the renormalization factor. The data collection, like the data initialization, can be done on paper,journal,subfield and field level.

Appendix B: Highest Cited Papers

R_c	R_I	Year	Title
1	1	1970	Cleavage Of Structural Proteins During Assembly Of Head Of Bacteriophage-T4
1	63	1971	The Assessment And Analysis Of Handedness: The Edinburgh Inventory
1	1	1972	Regression Models And Life-Tables
1	19	1973	Relationship Between Inhibition Constant (K1) And Concentration Of Inhibitor Which Causes 50 Per Cent Inhibition (I50) Of An Enzymatic-Reaction
1	1	1974	Film Detection Method For Tritium-Labeled Proteins And Nucleic-Acids In Polyacrylamide Gels
1	1	1975	Detection Of Specific Sequences Among Dna Fragments Separated By Gel-Electrophoresis
1	1	1976	Rapid And Sensitive Method For Quantitation Of Microgram Quantities Of Protein Utilizing Principle Of Protein-Dye Binding
1	1	1977	Dna Sequencing With Chain-Terminating Inhibitors
1	11	1978	Rapid Chromatographic Technique For Preparative Separations With Moderate Resolution
1	1	1979	Electrophoretic Transfer Of Proteins From Polyacrylamide Gels To Nitrocellulose Sheets - Procedure And Some Applications
1	6	1980	Ligand - A Versatile Computerized Approach For Characterization Of Ligand-Binding Systems
1	1	1981	Improved Patch-Clamp Techniques For High-Resolution Current Recording Fromcells And Cell-Free Membrane Patches
1	1	1982	A Simple Method For Displaying The Hydropathic Character Of A Protein
1	1	1983	A Technique For Radiolabeling Dna Restriction Endonuclease Fragments To High Specific Activity
1	2	1984	A Comprehensive Set Of Sequence-Analysis Programs For The Vax
1	4	1985	A New Generation Of Ca-2+ Indicators With Greatly Improved Fluorescence Properties
1	3	1986	Statistical Methods For Assessing Agreement Between Two Methods Of Clinical Measurement
1	1	1987	Single-Step Method Of Rna Isolation By Acid Guanidinium Thiocyanate Phenolchloroform Extraction
1	8	1988	Development Of The Colle-Salvetti Correlation-Energy Formula Into A Functional Of The Electron-Density
1	32	1989	Gaussian-Basis Sets For Use In Correlated Molecular Calculations .1. The Atoms Boron Through Neon And Hydrogen
1	2	1990	Basic Local Alignment Search Tool
1	2	1991	Molscript - A Program To Produce Both Detailed And Schematic Plots Of Protein Structures
1	5	1992	The Mos 36-Item Short-Form Health Survey (Sf-36) .1. Conceptual-Framework And Item Selection
1	1	1993	Density-Functional Thermochemistry .3. The Role Of Exact Exchange
1	1	1994	Clustal-W - Improving The Sensitivity Of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties And Weight Matrix Choice
1	4	1995	Genepop (Version-1.2) - Population-Genetics Software For Exact Tests And Ecumenicism
1	2	1996	Generalized Gradient Approximation Made Simple
1	1	1997	Gapped Blast & Psi-Blast: A New Generation Of Protein Database Search Programs
1	2	1998	Crystallography And Nmr System: A New Software Suite For Macromolecular Structure Determination
1	2	1999	Mechanisms Of Disease - Atherosclerosis - An Inflammatory Disease

TABLE I. Publications with highest *cRank* for each year.

[1] D. J. de Solla Price, Science **149**, 510 (1965), http://science.sciencemag.org/content/149/3683/510.full.pdf.

[2] M. L. Wallace, V. Larivire, and Y. Gingras, Journal of Informetrics **3**, 296 (2009).

[3] Redner, S., Physics Today **58**, 49.

[4] F. Radicchi, S. Fortunato, and C. Castellano, Proceedings of the National Academy of Sciences **105**, 17268 (2008), http://www.pnas.org/content/105/45/17268.full.pdf.

[5] M. E. J. Newman, Proceedings of the Na-

tional Academy of Sciences **98**, 404 (2001), http://www.pnas.org/content/98/2/404.full.pdf.
[6] A. Barabsi, H. Jeong, Z. Nda, E. Ravasz, A. Schubert, and T. Vicsek, Physica A: Statistical Mechanics and its Applications **311**, 590 (2002).
[7] R. K. Pan, K. Kaski, and S. Fortunato, Scientific Reports **2** (2012), 10.1038/srep00902.
[8] F. Havemann, M. Heinz, and H. Kretschmer, Journal of Biomedical Discovery and Collaboration **1**, 6 (2006).
[9] B. F. Jones, S. Wuchty, and B. Uzzi, Science **322**, 1259 (2008).
[10] R. Sinatra, P. Deville, M. Szell, D. Wang, and A.-L. Barabási, Nature Physics **11**, 791 (2015).
[11] M. Rosvall and C. T. Bergstrom, Proceedings of the National Academy of Sciences **105**, 1118 (2008), http://www.pnas.org/content/105/4/1118.full.pdf.
[12] M. Herrera, D. C. Roberts, and N. Gulbahce, PLOS ONE **5**, 1 (2010).
[13] C. Hurt, Information Processing & Management **23**, 1 (1987).
[14] L. Bornmann and H.-D. Daniel, Journal of Documentation **64**, 45 (2008).
[15] J. E. Hirsch, Proceedings of the National Academy of Sciences of the United States of America **102**, 16569 (2005).
[16] E. Garfield, Canadian Medical Association Journal **161**, 979 (1999), 10551195.
[17] O. Penner, R. K. Pan, A. M. Petersen, K. Kaski, and S. Fortunato, Scientific Reports **3** (2013), 10.1038/srep03052.
[18] R. Adler, J. Ewing, and P. Taylor, Statistical Science **24**, 1 (2009).
[19] S. Alonso, F. Cabrerizo, E. Herrera-Viedma, and F. Herrera, Journal of Informetrics **3**, 273 (2009).
[20] M. Bras-Amors, J. Domingo-Ferrer, and V. Torra, Journal of Informetrics **5**, 248 (2011).
[21] P. D. Batista, M. G. Campiteli, and O. Kinouchi, Scientometrics **68**, 179 (2006), http://www.akademiai.com/doi/pdf/10.1007/s11192-006-0090-4.
[22] T. Braun, W. Glänzel, and A. Schubert, Scientometrics **69**, 169 (2006).
[23] L. Egghe, Scientometrics **69**, 131 (2006).
[24] R. K. Merton, American Sociological Review **22**, 635 (1957).
[25] P. Chen, H. Xie, S. Maslov, and S. Redner, Journal of Informetrics **1**, 8 (2007).
[26] F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani, Phys. Rev. E **80**, 056103 (2009).
[27] D. F. Klosik and S. Bornholdt, PLOS ONE **9**, 1 (2014).
[28] P. D. B. Parolo, R. K. Pan, R. Ghosh, B. A. Huberman, K. Kaski, and S. Fortunato, Journal of Informetrics **9**, 734 (2015).
[29] T. Kuhn, M. Perc, and D. Helbing, Phys. Rev. X **4**, 041036 (2014).
[30] V. Larivière, É. Archambault, and Y. Gingras, Journal of the American Society for Information Science and Technology **59**, 288 (2007).
[31] This assumption is for simplicity, but it is grounded in the reality: The number of citations an article receives spikes few years after its publication [28].
[32] R. J. Glauber, Phys. Rev. Lett. **10**, 84 (1963).
[33] Physics Today (2014), 10.1063/pt.5.2012.
[34] R. K. Pan, S. Sinha, K. Kaski, and J. Saramäki, Scientific Reports **2** (2012), 10.1038/srep00551.
[35] A. L. Porter and I. Rafols, Scientometrics **81**, 719 (2009).
[36] E. Garfield, Science **122**, 108 (1955).
[37] A. F. J. van Raan, Scientometrics **62**, 133 (2005).
[38] K. W. Boyack and K. Börner, Journal of the American Society for Information Science and Technology **54**, 447 (2003).
[39] D. W. Aksnes, Journal of the American Society for Information Science and Technology **57**, 169 (2005).
[40] P. O. Seglen, BMJ **314**, 497 (1997).
[41] E. Favaloro, Seminars in Thrombosis and Hemostasis **34**, 007 (2008).
[42] B. S. Frey and K. Rost, Journal of Applied Economics **13**, 1 (2010).
[43] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Rev. Mod. Phys. **87**, 925 (2015).
[44] C. Castellano, S. Fortunato, and V. Loreto, Rev. Mod. Phys. **81**, 591 (2009).