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

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

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

Keywords science of science, scientometrics

ISBN (printed)	**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

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed

gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Espoo, June 2, 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 Pietro della Briotta Parolo, Raj Pan,Francesco Becattini,Marija Mitrovic,Arnab Chatterjee,Santo Fortunato. 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 the Journal of Informetrics, January 2017.

IV Pietro Della Briotta Parolo, Mikko Kivelä, Kimmo Kaski, Santo Fortunato. On the Shoulders of Giants: Tracking the Cumulative Information Spreading in Citation Networks. *preprint*, Submitted to Phys. Rev. X, June 2017.

Author's Contribution

Publication I: "The Nobel Prize delay "

The author gathered part 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 Information Spreading in Citation Networks"

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

1. Introduction

The underlying driving force of science has always been the one 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 the one of the first ones to directly tackle the pattern of bibliographical references, but it also has 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 field of Network Theory was still in its earliest stages.

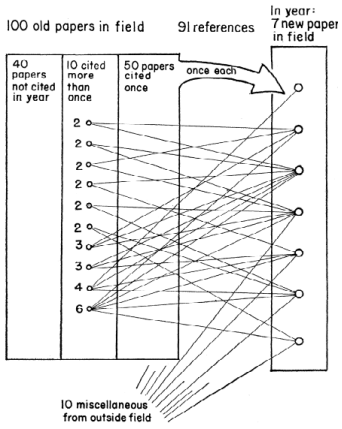

Figure 1.1. Representation of citations as a network structure

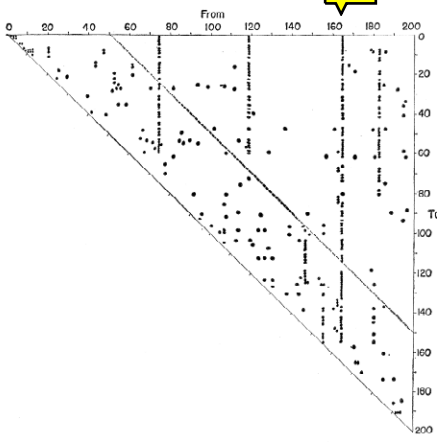

Figure 1.2. Representation of citations as an adjacency matrix.

What is most striking however, is that already in its origins, the study of the scientific production has required an analsyis 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 itself to rely one's work on the top of previous ones and therefore adding a temporal dimension to its devolopment, as new discoveries and breakthrough appear and link themseleves 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. 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 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 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.

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 1 talks about the study of citation patterns, with their properties, biases and attempts at modelling them. Chapter 2 introduces the basic concepts of network theory and how these concepts have been used to analyze the social and collaborative structure of science. Chapter 3 talks about the efforts in trying to determine the quality of scientific publications by the development of *metrics*.

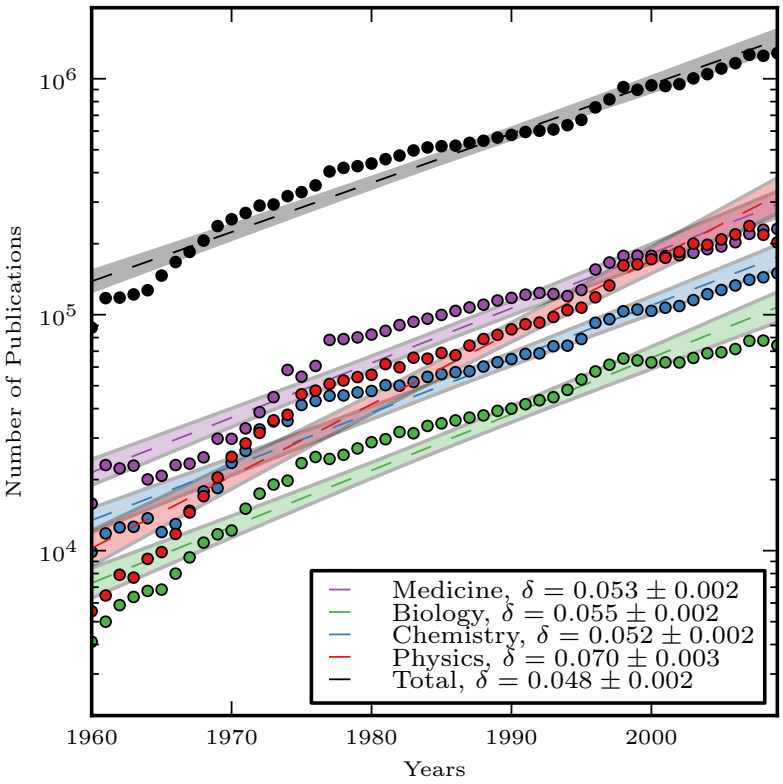

Figure 1.3. Growth of publications. From Publication I.

2. Citations and 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 aknowledge 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 Galielo 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. Futhermore, 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; while the results in aging literature is slowly assimilated as a basic result, scientists move on to newer results as the basis of their works, thus implicitly determining when a grounbreaking 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 quantiative evaluation of the success of a paper and is sufficient to be able to

compare and rank publications as well as scientists. However, the aforementiond process of obsolence in science adds a dimension which has been described as an *attention economy* [5] 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 relvant 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 modelling mathematically the citation patterns of scientists.

2.1 Citation Distributions

One of the eariest questions that scientometrics tried to answer already with Desolla's seminal paper has been: *What is the functional form of the distributions of citations?*. In particular, the interest was in the tail of the distributions, as the average number of citation is bounded to be low due to the finite number of references available, while a few exceptional papers are able 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 [6] 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 [7]. It is important to notice that such analysis was possible to carry 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 ISI and 24 thousand papers from Physical Review D) combining for more than 7 milion citations, the author was able, for the first time, to carry out a thourough computational statistical analysis of citation distributions. The results offered an interesting and, to a certain extent, worrysome insight of the relative popularity of scientific publications: almost half of the papers failed to gather any citation at all, with 80% of the publications gathering 10 citations or less. Even though also de Solla noticed a huge amount of uncited paper, Redner was able to confirm the pattern also for a much larger and significant data set. The author conclude that a final evaluation of the functional form of the citation distribution cannot be thouroughly computed as the tail of the distribution has not reached its final state, as the highly cited papers are still gathering citations. He also points 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 [8], providing evidence of a power law behaviour with $\alpha \approx 3$, compatible with de Solla's findings. Furthermore, the author conclude with what can be considered the *cookbook* for future attempts at modelling the citation mechanism: a short memory (or Myopia) and a "rich get richer" kind of mechanism. The latter would become a massive topic starting from the following year, with Barabási's work on scaling in random networks [9] 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 [10]. 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. [11] 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 are extremely different across fields. This is a well known bias, the roots of which lie in the different sizes of the fields or disciplines [12] as well as in different conceptual meaning of the citation itself [13]. In order to get rid of discipline dipendent 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 $\sigma^2 = \mu$ allows the expected average value of c_f to be 1, thus allowing to compare the distributions across disciplines. The authors also find that the collapsing behaviour persists also when distribution from different years are compared, therefore suggesting that the functional 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 it can be quantified and corrected. This can be done by comparing the relative values of citations across different by limiting the citation count to citations within the field and by doing a quantile-quantile analysis of this new measure c' versus the original citation count, thus determining how relevant a citation count is within each field [14]. Fig. 2.1 shows the quantile-quantile plot for a selected number of subfields. Data shows that for most of the subfields, there exists a shared scaling relation in form of a power law, characterized by parameters that are found to be stable across time, therefore confirming that citation distributions follow an underlying universal behaviour.

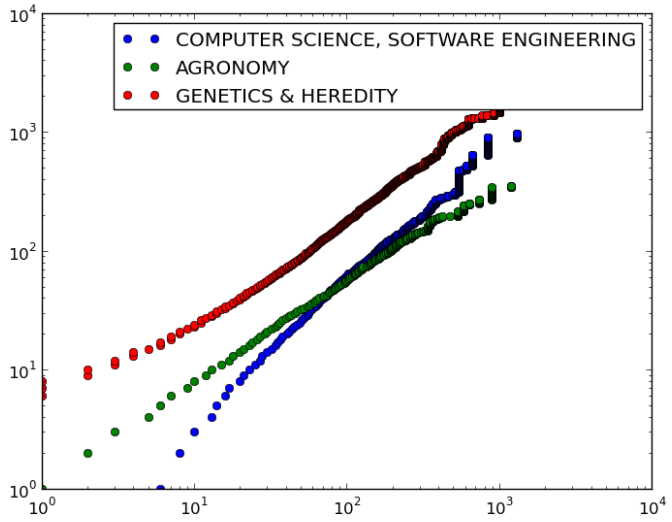

Figure 2.1. c' vs c from [14] reproduced with our data set.

2.2 Biases in Citations

In 2005 Hajra et al. were [15] 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 [16], 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 obsolence 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.

While the average suggests that papers are being forgotten within a limited period of time, other works looked 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 suddendly highly influential and cited [17]. 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 $t_m \in [0, T]$. 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

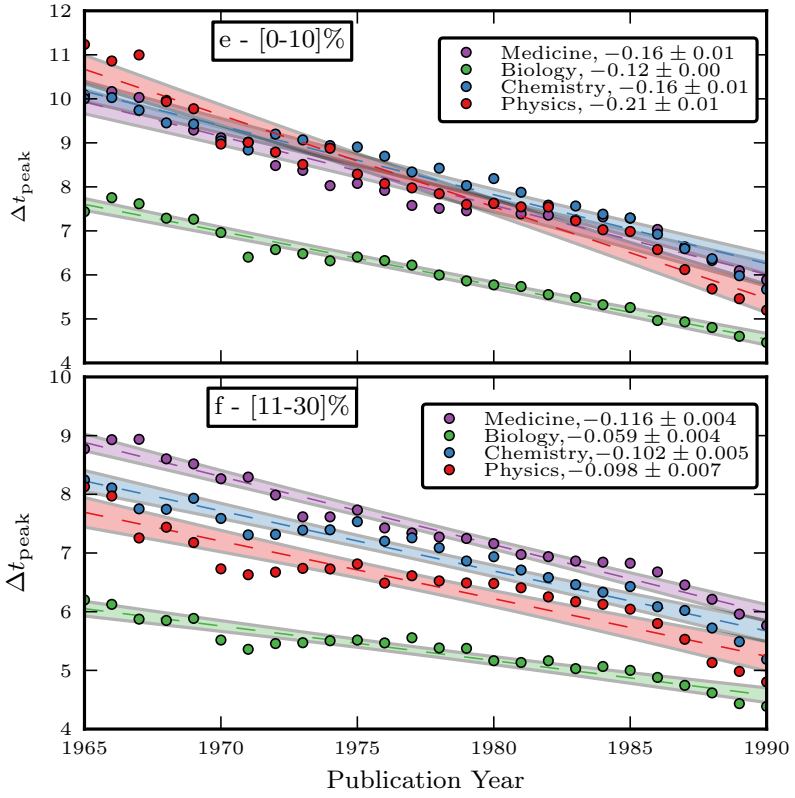

Figure 2.2. Time evolution of the mean values of time to peak Δt_{peak} for top 10% (top) and [11-30]% percentiles (bottom). The mean value $\langle \Delta t_{\mathrm{peak}} \rangle$ decreases linearly in time. The linear fit, 95% confidence interval and the slopes of the linear fits are also shown. Taken from Publication III.

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 example being the recent experimental discovery of the Higgs boson [18], the disovery of was confirmed only in 2012 thanks to the development of the LHC at CERN in Geneva[19]. 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 phenomenonoligcal studies regarding its discovery had already started [20], as one of these studies points out :

"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." [21] The temporal aspect of recognition of older theoretical brekthrough 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 increasingly at a very high rate, to the point where the original authors might pass away before seeing their discoveries empirically confirmed.

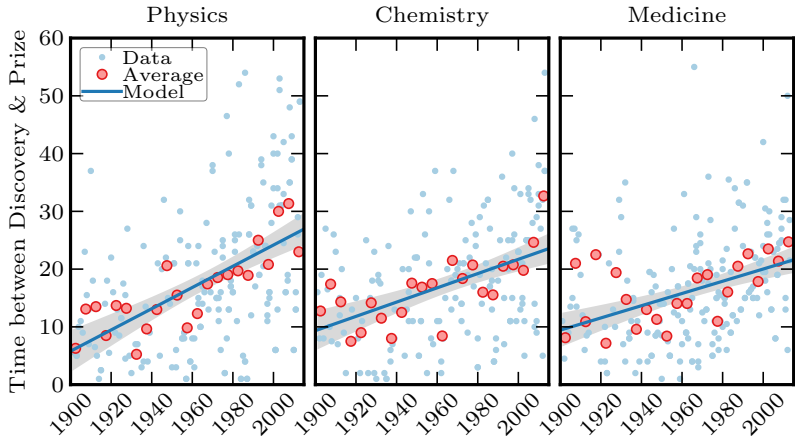

Figure 2.3. Time lag between discovery and nobel prize, created with data from Publication I.

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 to access 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 and read more recent material has consequently introduced a change in the way authors update their knowledge. The effects on the scientific comunity were rapid, as in 2003 already De Groote et al. [22] showed through a survey that general users of scientific material prefer the digital copies to the printed ones. The constant need for

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 [23]. An interesting study in the impact of online available material on citation patterns came in 2008 when Evans[24] reported that the rise of online available publications shifted the citation patterns by studying the effect of online availability of journal issues within the citation patterns of the journals. The results showed that as 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 hyperlink, 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. [25] as well as by Pan et al. [26]. Verstak used Google Scholar Data to analyze all publications available between 1990 and 2013. The authors calculated the fraction of references in these papers point at least 10 years before the year of publication for each paper and found that such fraction is actually *increasing* in time. Furthermore, they notice that the value of the change over the second half of the period studied was much larger than in the first, with these two periods with the former matching the period in which digitalization has took place (2001-2013), therefore concluding 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 deviced 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 *redirection* to other articles cited by papers 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 metally keep track of all relevant publications being published might be among one of the greatest limitating factors in citation patterns, as [27] reports 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 show 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 endavour,

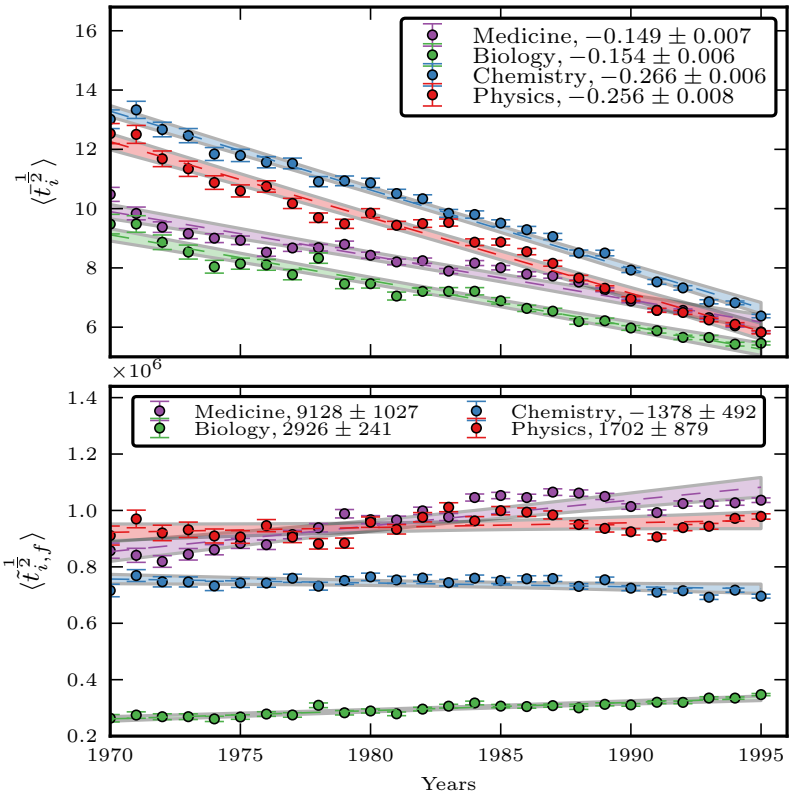

Figure 2.4. The evolution of the half life of papers after the peak $\langle t_i^{\frac{1}{2}} \rangle$ in terms of absolute time (top) and $\langle \tilde{t}_{i,f}^{\frac{1}{2}} \rangle$ in terms of the number of publications (bottom) for the four different fields and for the top 10% percentile. The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

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 [28] researches similar work are more likey to cite one another, surpassing friendhsip as a predictor of reciprocal citation. Similarly, [29] 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[30], with the number of authors not being a factor though. However, the authors point out that the most dominating contribtion 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 carreers 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"* [31], 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 notoriety, hoping for success to accumulate from there. However, while self-citation does show to have an impact on citation counts, it is not clear whether the correlation is linked to a matter of *visibility*, i.e. trying to put forward one's results as a "bandwagon" effect, or rather a matter of *qualtiy*, as one author mentions its own best works as a basis for future ones [32]. More recent results confirm [33] that the trend is still significant, yet retaining different patterns in different fields, due to the possibiliy 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-citations leads to a higher chance of inter-citations of the second order, with collaborators of collaborators being more likely to be cited. An author might also influence its own carreer retroactively as shown by Mazloumian et al. [34]. The authors found out that grounbreaking results by an author have a positive impact on the previous literature by the same author, 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 instrinsic 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 [35]. This physcosociological mechanism that enhances the carreer 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 [36]. In general, a citation bias towards successful papers (preferential attachment) and one towards successful authors

(Matthew Effect) show 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 [37], which I will not analyze for the sake of brevity.

2.3 Modelling

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 heterogenous results within the same field of analysis. It is therefore not surprising that the pursue for a mathematical model that could correctly reproduce the properties of citatin 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 in the temporal aspect of the models.

The earliest and most sucessful attempts at modelling citation dynamics lie in the *rich get richer* or, technically speaking, *prefential 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[9]. In his work, Barabasi suggests that a mechanism 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: $A(c) \propto c$. is able to explain the citation distribution both from a qualitative point of view (its power law 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 complexy systems, with particular success in biology [38, 39], of which citation dynamics represent one of the examples.

A confermation of the preferential attachment came in 2005 with Redner [10], in which the author d to confirm 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 modelling framework. While theoretical works tried to tackle the topic from a purely mathematical standpoint [40, 41],

it was Hajra et al. [16] in 2006 who applied it with success to the modelling of citations. The authors followed the previous theoretical works and formulated an for the attachment rate of $\Pi(c,t) = C(c)T(t)$, trying to identify the functional form for the temporal aspect that would best fit the data through the analysis of the distribution of citation ages $Q(t)$. 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 indentified 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 $t_c \sim O(10)$ 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 hyphotetized 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 hierarchicy 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 frameowrk is known to be able to explain the evolution of the response to new pieces of information online [42], 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 model came in 2008 with a work by Wang et al. [43]. Their model proposes to not separate globally the dependence of the attachment rate to the two variables, considering the aging process to be related not to the whole paper, but to the citations themseleves. The logic behind this idea is that a paper who 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 somwhat 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 modelling part only to reproduce only certain aspects of the citation dynamics with still a focus on the temporal aspect. In 2001 Burrel was able to confirm that in presence of a stochastic process that assigns citations to publications based a non-homogeneous Poisson process[44] is bound to produce articles that will remain citedless. In 2009 Wallace et al[45] 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 hypothetized 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 Y, with citations being given randomly through a Poissonian distribution, given the size of the two variables. The authors then limit the probablity 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. It therefore follows 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 [46] 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 thefore 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 show that if on average newly published papers cite m earlier papers, chosen proportionally to the number of citations k they already have, plus a variable r, it is normal to expect more recent papers to gather constant r. From this model one can calculates 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 with $t \to 0$, i.e. more recent papers These results are somwhat in contrast with the previous discussion regarding obsolesence and the timespan 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 [47] 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 humany dynamics [48, 49]. 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. $\Delta c/c = [c(t+\delta t)^i_{in} - c(t)^i_{in}]/c(t)^i_{in}]$,

where $c(t)^i_{in}$ is the number of incoming citations a paper received at time t. 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 suddendly receive orders of mangitude of citations more than they ever did, especially during the early years of a paper. 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 dependance 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 probablity of a certain paper j to be cited described as : $\Pi(i \rightarrow j, t) \propto [c^j + A_j(t)]$. 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 citajon 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 probale (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 unsuccess might be sufficient to prevent a high quality paper from rising to notoriety. In 2015 Wang et al. [50], 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 agin (i.e. it's not the fitness that decays, but rather the *novelty*) comes at a cost, as the authors needed to introduce to 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 themseles and be quantified through a least square fit method. Furthermore, the author managed to quantify the importance of the individual contributions within the attachment rate formula, finding that the dependance 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 paper begings with these two sentences: *"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 [51]. In general, the intrinsic collaborative nature of science either by cumulative contribution (the shoulder 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.

Network theory showed for the first time its potential to investigate 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 city exactly once? Until the midlle of the 19th century, network theory remaiend confined as a branch of topology in theoretical mathematics, with a few cases of applications to other fields. However, in the post war period sociologists understood that a matrix based representation, i.e. the underlying bedrock of network theory, of social ties could be beneficial for the study of social structures [52, 53]. The breakthrough came in 1959 with Erdös and Rényi's work on random graphs [54] 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 [55] and was later used as a fundamental tool for studies that required a network based structure, in particular for models in epidemiology

[56, 57]. The model, despite being extremely versatile in its semplicity, failed to reproduce charachteristics of large scale networks that the digitalization of society allowed to gather and analyze in the 90's. This required the development of new models, which rapidly happened [58, 9].

Since then network theory has managed to be applied in a large spectrum of fields, dealing with non-trivial network structures that required ad-hoc methods and algorithms, leading to a whole new field, often referred to as *complex networks*, in order to differentiate it from the purely mathematical one. As the theory developped, the application of its methods to publication data became a fertile branch of the field. This Chapter will first go through the basics of networky 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 $G = (P, E)$ where P is a set of N nodes and E is a set of M links (or edges) connecting pairs of nodes. A simpler way to represent a network is through its *adjacency matrix* A, which fully describes the graph. Its elements a_{ij} can be 1 if there is a link connecting node i and nodej and 0 otherwise. If A is symmetric the graph is undirected as all of its links go in both directions. If $a_{ii} \neq 0$ self loops are considered. The elements of the matrix are usually binary and symmetric, thus only indicating whether two nodes have a connection or not. *Directed graphs* take into accoun the directionality of the links by dropping the symmetry requirement. Similarly, weighed 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 [59]. 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 can represent a problem computationally in terms of storage space as it requires to store N^2 entries, most of which do not carry information. Since it is not rare for matrices to be sparse [60], adjacency matrices can be replaced with *adjacency lists* in which each row i enumerates the neighbours of the node along with the value of the edge in case it is required. Recently, there has also been a necessity to analyze networks temporally [61] in order to take into account the changes in the activity of edges.

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 $k_i = \sum_j 1$ if $a_{ij} \neq 0$,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 necesarry to analyze them statistically. In the ER model [1] each link exists with probability $p = \frac{M}{N}$, leading the probability of node i to have degree k as the probability of having k times successful binomial extractions, thus converging to a Poissonian distributions as the size of the network grows, with k remaining constant. However, empirical evidence [63] has shown that real world networks have a dramatically different behaviour when it comes 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 be scale-free, i.e. with the existance of nodes with large degree called *hubs*, along with a vast amount of nodes with low degree values, with the distribution that follows a power law behaviour $P(k) \propto k^{-\alpha}$. In 1999, Barabási and Albert proposed a different model, in which each 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.

Another fundamental properties 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 [64, 65] 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 can be 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 PAM 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 PAM model the evidence is less trivial, but it comes from the fact that hubs have a tendency to get citations from all new nodes, thus failing to select connections to specific nodes. Interestingly, real life networks show different scenario, with certain networks

[1] This formulation was presented in the same year by Gilbert [62] and is statistically equivalent to the ER model.

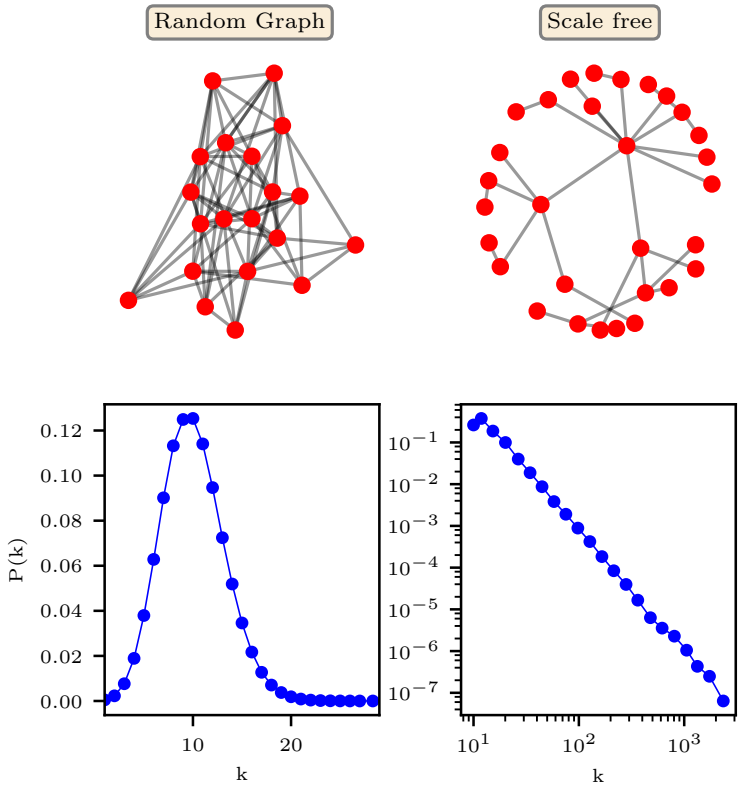

Figure 3.1. Difference in topology and degree distribution between a random graph (left) and a scale-free network (right).

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 [66]. 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 [64] show that a high assortativity links 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, being 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 [67] and have therefore been often subject of systematic failures [68] due to their structural fallacies.

3.1.2 Clustering, paths and distances

The Clustering Coefficient measures how likely two nodes within the neighbourhood of a node are likely to also be connected:

$$C_i = \frac{|e_{jk} : v_j, v_k \in N_i, e_{jk} \in E|}{\frac{k_i(k_i-1)}{2}} \tag{3.1}$$

The denominator represents the maximum number of possible links between common neighbours as it is the number of diagonals in a polygon of k_i vertices. The numerator instead is a measure of how many couples of neighbours are linked temselves. In case of a directed graph, the simmetry is broken and therefore the coefficent 2 disappears from the definition. 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) form a triangle and can be applied to both undirected and directed networks In an undirected network the mean distance 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 fully connected (i.e. there are parts of the networks that are separated), the value of the distance 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.

In 1998 Watts and Strogatz published a paper that showed how the currently available models based on random graphs were unable to grasp the properties of real networks in terms of clustering coefficient and path length[58]. While their analysis of diverse networks (power grids, biological networks, film actors) showed large CC and short paths, the ER model [69] is bound to generate networks with average path length $\propto log(N)$ and have an extremely low value for the CC. They proposed a 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 decreases from 1, local clustering remains high while paths between distant nodes cause a significant reduction of the average path lenghts. They called their networks *small world networks* in reference to the famous social experiemnt of the six degrees of separation [70], which was the first attempt at calculating path lenghts in social networks. Small world networks thus managed to explain key features of social networks that not even the PAM managed to outperform. Similarly, they managed to find a theoretical validation of the importance of weak ties connecting strongly connected groups formulated by Granovetter in the 60s [71].

3.1.3 Communities and modularity

Between 1970 and 1972 Wayne W. Zachary collected data about the interaction between 78 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 the club with one instructor and the other half leaving it [72]. 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. 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 [73], urban areas and societies can be structured in large groups divided by language [74], and also network scientists are organized in communities [75]. While communities are easy to qualitatively define, their mathematical deifnition 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.

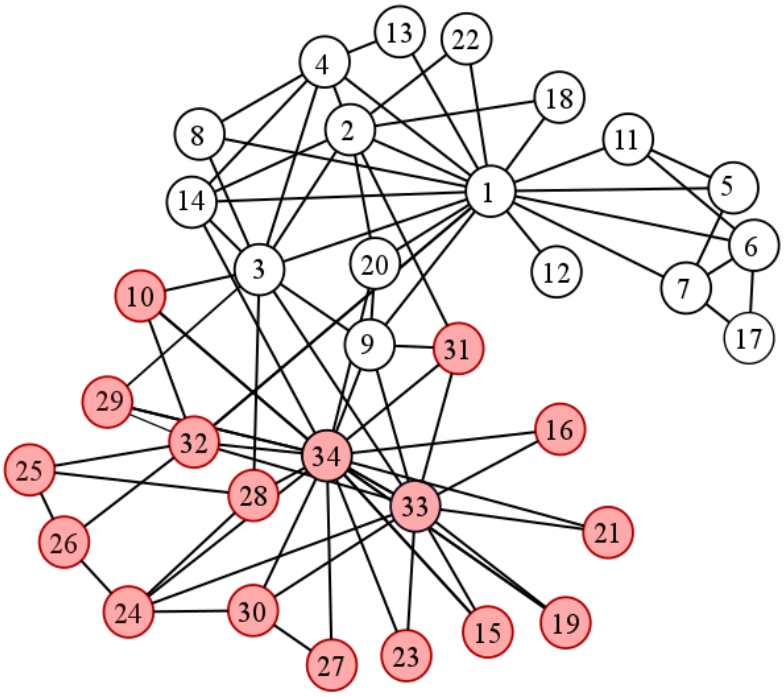

Figure 3.2. Karate Network [72].

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* [76]. This method,

introduced by Girvan and Newman in 2002 [77] is based on the idea that a good partitioning maximizes the amount of edges within a community and minimizes the amount of links towards the outside of the community. 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. We calculated for new nodes joining the EN the fraction of references that stay within the EN, thus quantifying how modular the EN is. We showed that the EN has a sharp initial in modularity that saturates within 10 years, before gradually decreasing. Unfortunately, despite its simplicity, modularity also offers some limitations. Fortunato et al. in 2007 showed that modularity optmization is bound to have a resolution limit, i.e. a minimum size of communities under which the method fails to detect communities [78], which can represent an issue as real networks can be organized in hierarchical or tree-like structures [79]. 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 [80]. Furthermore, another key limitation is the presence of multiple suboptimal solutions [81] 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 non-rigid definition of a community, thus forcing the scientists to perform a trial and error analysis based on the cumulative information gathered in the process [82].

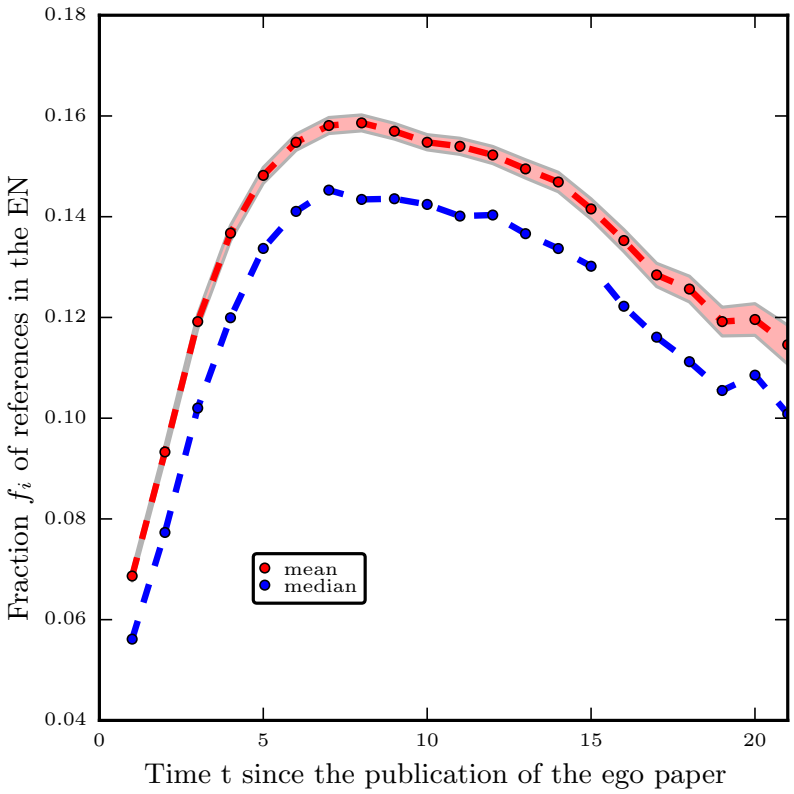

Figure 3.3. Time evolution of the mean and median of the fraction f_i of references of papers of the full Ego Network belonging to the Ego Network as a function of the number of years since publication from Publication III.

3.2 Author networks

As we have seen, network theory provides a solid framework with which to investigate social structures. It followed therefore that scientstists 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 scientsists 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[83]. 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 interestring tendency for authors to cluster, even though this might be biased by the presence of papers written by 3 or more authors, which create immediately a triangle in the network. Newman's work showed the intrinsic social nature of science as a network of collaborating node, with a community-based structure that is coherent with a preferential attachment method 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. [84], 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 evoltion of the individual carreers of the different authors [85]. Sociological considerations [86] can support the hypothesis that preferential attachment method is the the driving force only of collaboration only for scientists in the middle of the carreer (thus also in the middle of the distribution), with the tails of the distribution being dominated by either established scientist, who don't require to build up their newtork 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 evolution over time as *networks change over time, both because people enter and leave the professions they represent and because practices of scientific collaboration and publishing change* [87]. 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 [88]. 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 collabroative 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 network [89]. While 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 scients usually are not likely to cover long distances, and can play a crucial role in one's career [90]. Similarly, the choices of collaborators are also affected by geographical considerations that can be linked to policemaking from individual countries or unions [91, 92].

3.2.1 Ties and Carreers

In such framework, it becomes therefore fundamental to investigate the different nature of the connections that connect different authours at different stages of their carreers; afterall 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 carreers [93, 94], thus affecting the structure of collaborations both locally and in time. Similarly, as the network structures are known to influence team-performance [95, 96], it is natural to conjecture that these kind of mechanisms are reflected in the data of scientific collaborations. In order to do so 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 weighed 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 [59], collaborative networks show a unique charachteristic: dense neighbourhoods are the core structure of dense neighbourhoods, with strong links connecthing different neighourhoods. This effect is considered to reflect the hierarchical and temporal dimension of scientific carreers: as senior researches build strong ties with each other over

time, they form research groups composed of young resarchers [97, 98]. Even though it is only a few strong links between senior scientists that keeps the scientific network of authors toghether, 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 [97], which is hierarchically and sometimes inequally structured in its hiring system [99] and in which early developments can lead to a cumulative advantage in a career [36, 100] it appears evident that the evolution of the social and collaborative strucure 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 [101]. 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 instrinsic 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 [102, 103] of basic ideas and algorithms originally introduced decades earlier in the early years of quantitative sociological studies of social networks [104, 105]. The most common type of centrality is betweenness centrality [104], 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 [106]. Let 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 costant. 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 [107]

$$\lambda x = xA$$

and solving the related eigenvector equation. Eigenvector centrality can come in many forms [105] and is also the main idea behing Google's PageRank algorithm [108]. Practically, eigenvector centrality measures how central a node is by quantifying how many paths go through a certain node, ignoring whether they are the shortest paths or not. 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 [109].

Data shows that the values of centrality in co-authorship networks are extremely skewed, with scientists with the highest score being well separeted from the 2nd tier, which in turn is well separated from the 3d and so on, thus confirming the hierarchical structure of science [110]. Also, the weighed network analysis show 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 [111]. 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 [112, 113].

3.3 Paper 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 and it has been extensively studied in social contexts [114, 115]. In a social EN the new nodes that attach to the original ones 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 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 oter. This peaks after a few years, with the building of a strongly connected core while, however, islands of isolated papers start to appear and eventuall, 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.

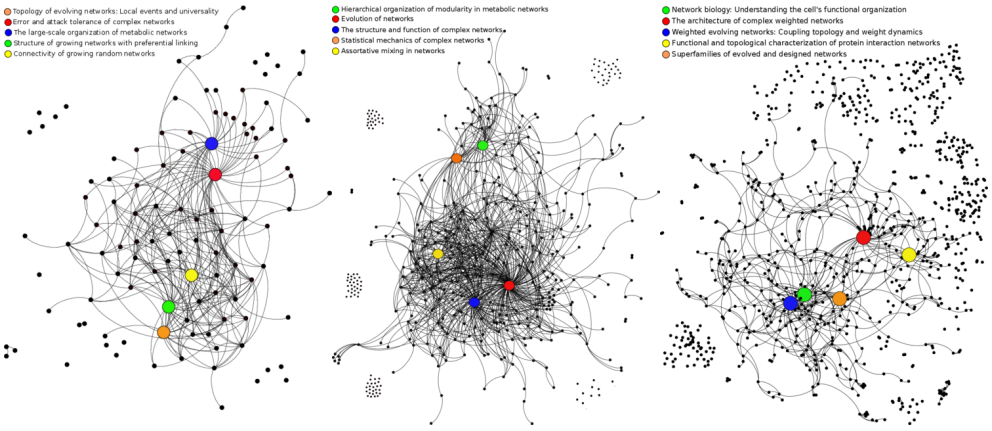

Figure 3.4. Ego-network for Barabási and R. Albert's paper on scale-free networks [9]. We consider windows of size $w = 2$ at t=1 (left), t=3 (center) and t=5 (right), where t is the number of years from publication. Therefore the windows are non-overlapping and cover the intervals 1-2, 3-4 and 5-6 (years 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. Taken 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. While 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 cirlces and thus allowing to quantify the size and shape of the *wake* of a paper [116]. 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 δ that 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 disoveries is not necessarily correlated to the first approximation, i.e. the citation count.

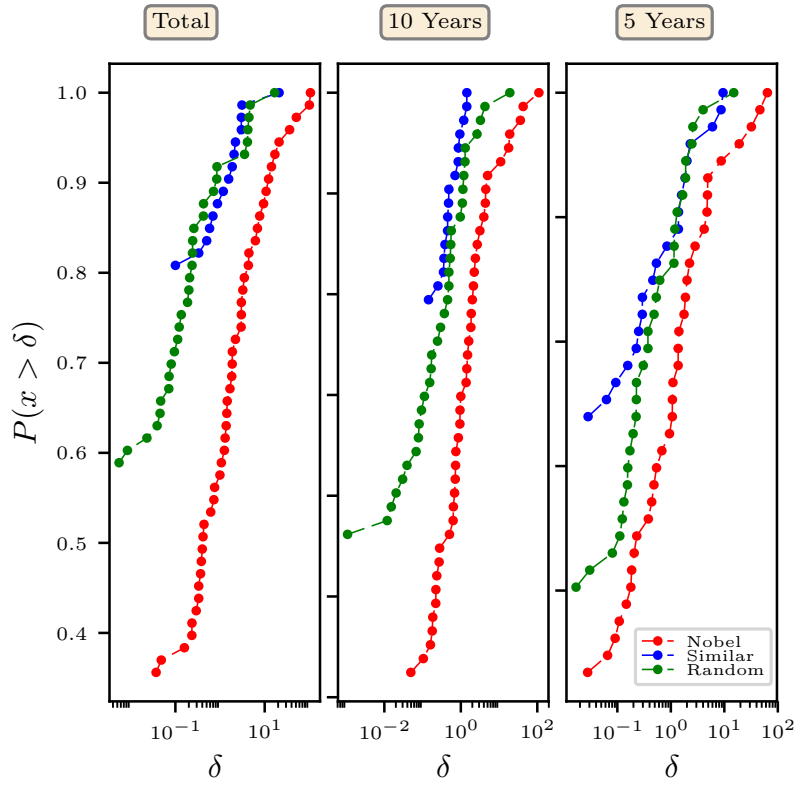

Figure 3.5. Cumulative distribution of δ for nobel papers, paper within a 3% in citation volume in the same time interval compared to nobel papers and for random papers. Only papers with positive δs are included. Taken from Publication IV.

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 [117]. 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 [118] is well known to be a driving force of the spread of new ideas [119] 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 [120]. 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* [121]. 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 [122]. 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 [123]. A meme based approach to the spreading of scientific ideas has been attempted with success [124], 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 classifical division of science in major branches such as Physics, Mathematics, Biology, Economics etc. is commonly used also oustide 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 simplicistic and efficient on a superficial scale, but we know science to be a intrinsically dynamic world. Bibliometric studies [125] and studies on the co-occurance network of scientific terms [126] has shown that fields themselve 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. A SEIR epidemic model in which scientists transition from being Susceptible to a new idea (i.e. working in a related field) to being Exposed to it (i.e. they have found out about it) and then proceed to become Infected, thus spreading the idea before ultimately Retiring. Emprical evidence shows that the population growth of fields can be modelled with sucess by this model [127]. 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 [128], in which he described the process by which scientific knowledge progresses as being composed of periods of staticity separated by abrubpt chanes 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 exaples 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 coeherence 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 pehonomena or the publication of new innovative work which ulitmately leads to completely new fields being born in a relative short time. Therefore one can 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 prosess is accompanied by a topological transformation in the structure of the author network of the field: intially 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 insuccess of the original idea [129]. 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 [130]. 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 [131], Furthermore, fields themselves are not monolithic blocks, but rather can be organized in structured hierarchical layers; Physics for example, manifests in its own paper netwrk a number of subfields that have different local structure, with smaller subfields being more self-referential and thus more modular [132]. 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 [133]. 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 [134, 132]. Multidisciplainarity is slowly increasing and it can be analyzed in terms of the flow of information across fields [135], a technique that has led to the possibility of determining the stabilization of interdisciplinary fields thus becoming new stand alone disciplines [136]. 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's 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, it is the one who is slowing down the most in its tendency to share information, probably as a consequence of it growing to the level of a stand-alone discipine with increased levels of self-referentiality.

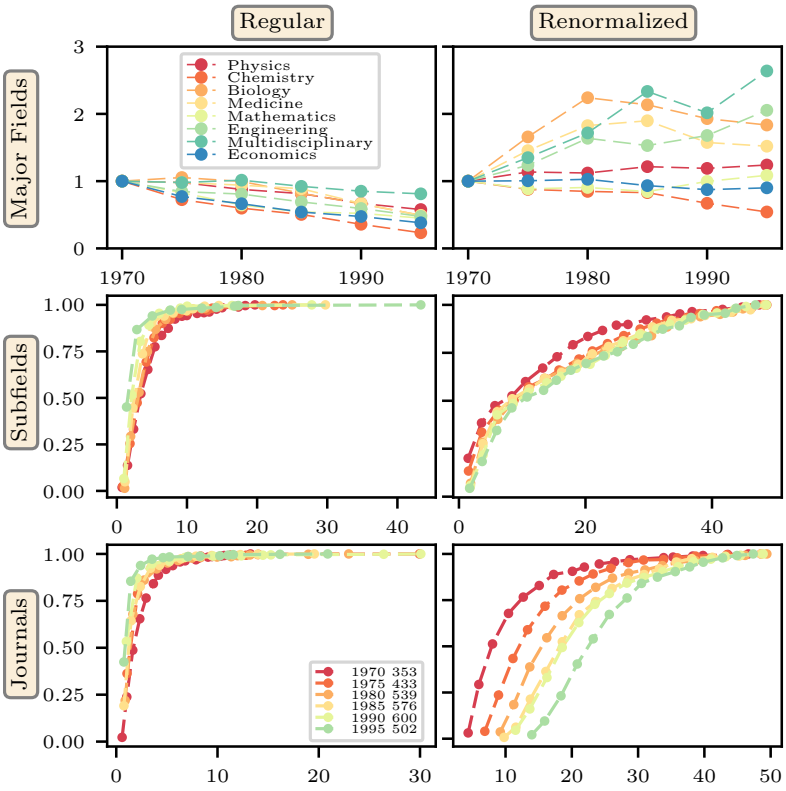

Figure 3.6. Changes in half life in time for the regular (left column) and renormalized scneario (right) and for different grouping of papers. Taken from Publication IV.

4. Science and Metrics

Here I dicuss the need for quantitative evaluation of scientific outputs, which lead to the development of metrics. 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 reasearch 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 [137] or, in general, influencing the structure of Academia itself [138], thus compromising the previous underlying assumptions of citations as a free and voluntary choice.

[139][140] among first papers to warn amout using citation metrics for evaluation of quality.

4.1 Paper Rankings

4.2 Author based Rankings

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

[5] 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.

[6] 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.

[7] 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.

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

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

[10] Redner, S., “Citation statistics from 110 years of physical review,” *Physics Today*, vol. 58, p. 49.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

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

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

[22] 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.

[23] 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.

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

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

[33] 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.

[34] 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.

[35] 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.

[36] R. K. Merton, "Thein Science," *Science*, vol. 159, pp. 56–63, Jan. 1968.

[37] 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.

[38] 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.

[39] 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.

[40] 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].

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

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

[43] 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.

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

[45] 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.

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

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

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

[49] 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.

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

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

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

[53] 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.

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

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

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

[57] 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.

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

[59] 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.

[60] 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.

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

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

[63] 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.

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

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

[66] 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.

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

[68] 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.

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

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

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

[72] 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.

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

[74] 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.

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

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

[77] 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.

[78] 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.

[79] 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.

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

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

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] "Nstudies in scientometrics. part 1. tran- ¨ sience and continuance in scientific authorship.," *International Forum on Information and Documentation*, p. 17–24, 1976.

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

[88] S. Uddin, L. Hossain, and K. Rasmussen, "Network effects on scientific collaborations," *PLOS ONE*, vol. 8, pp. 1–12, 02 2013.

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

[90] 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.

[91] 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.

[92] 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.

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

[94] 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.

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

[96] P. Alex, "The new science of building great teams.," *Harv Bus Rev*, vol. 90, pp. 60–69, 2012.

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

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

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

[100] 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.

[101] 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.

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

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

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

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

[106] 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.

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

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

[109] 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].

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

[111] 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.

[112] 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.

[113] 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.

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

[115] 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.

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

[117] 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.

[118] 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.

[119] 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.

[120] 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.

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

[122] 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.

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

[124] 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.

[125] 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.

[126] 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.

[127] 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.

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

[129] 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.

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

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

[132] 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.

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

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

[135] 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.

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

[137] 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.

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

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

[140] 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.

Publication I

Pietro della Briotta Parolo, Raj Pan,Francesco Becattini,Marija Mitrovic,Arnab Chatterjee,Santo Fortunato. 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

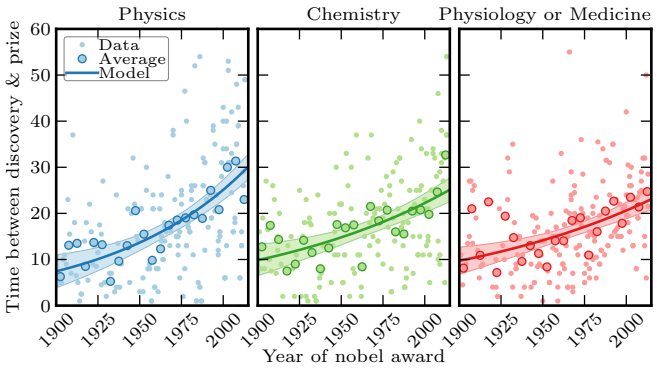

FIG. 1. Time difference (in years) between the discovery and the awarding of the Nobel prize, versus the year when the award is received. Each plot shows the raw data, the 5-year average, and the exponential fit with its confidence interval. The lag is increasing for the three fields, with rates of 0.012±0.002,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

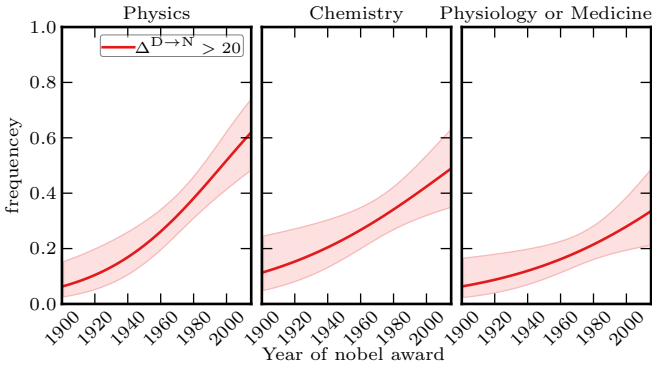

FIG. 2. The frequency of prizes awarded over 20 years of the discovery is increasing for all disciplines. The growth is fastest for Physics and slowest for Physiology or Medicine.

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

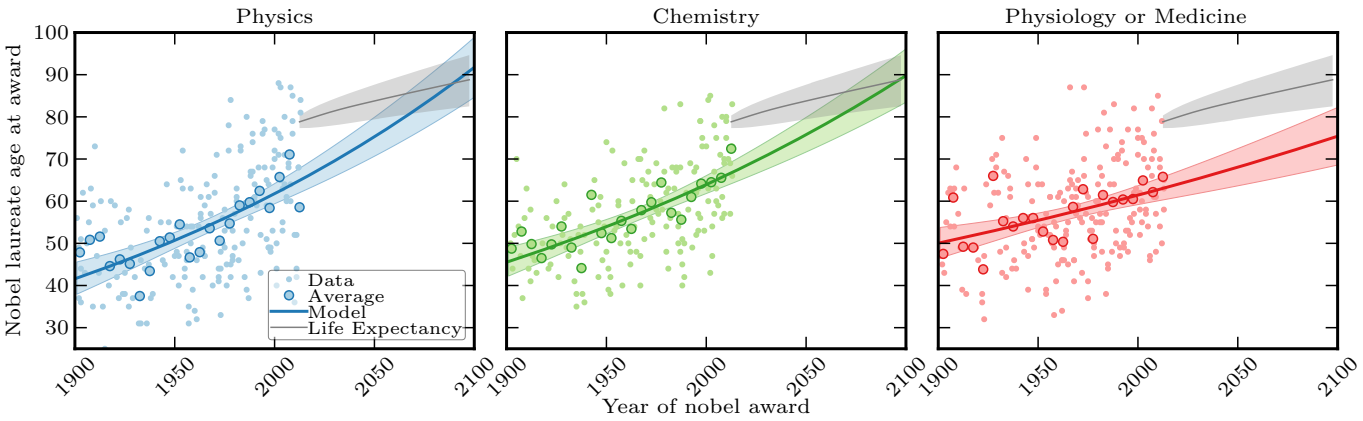

FIG. 3. Change in the age of the scientist at which the Nobel prize is awarded. For all fields there is an increasing trend. For Physics and Chemistry the rate of increase is similar (0.0040 ± 0.0005 and 0.0034 ± 0.0004), while for Physiology or Medicine the increase is much smaller (0.0020 ± 0.0005).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

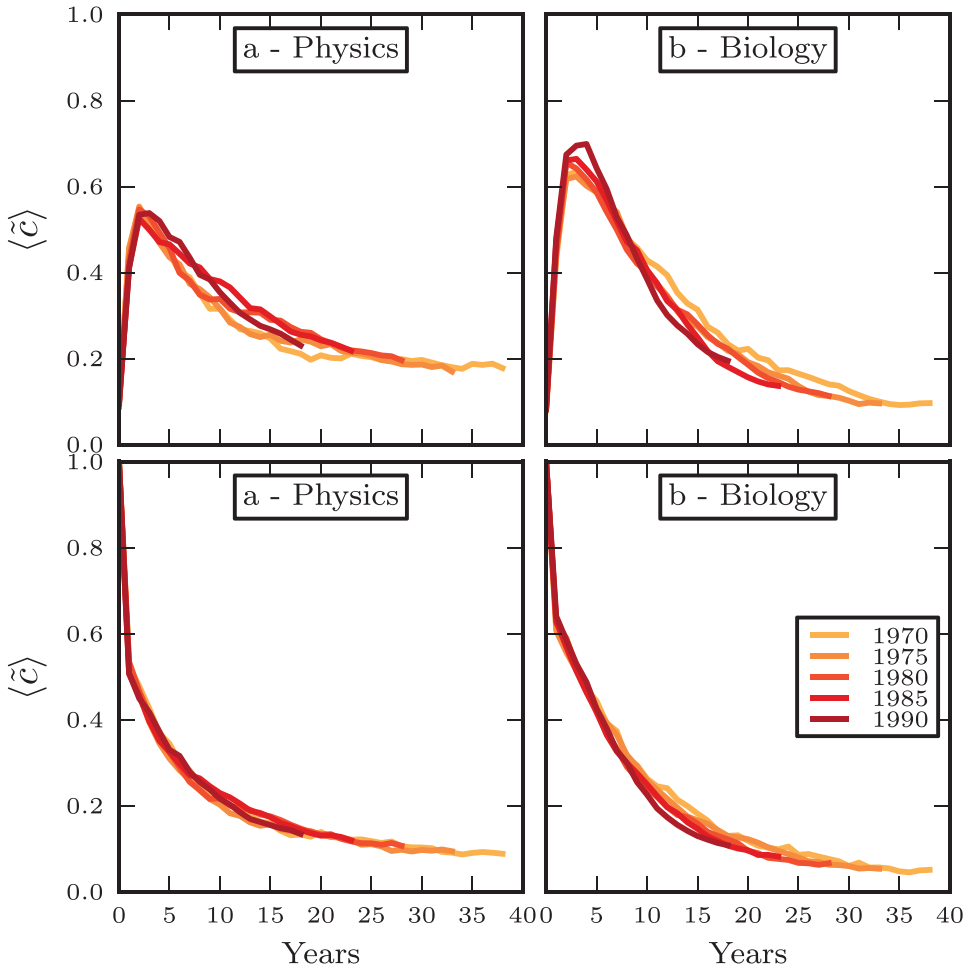

Fig. 1. The citation life-cycle is both field dependent and time dependent. (Top) Normalized number of citations per year received by papers in Physics and Biology published in the same year, for different publication years. Normalization is done by dividing the number of citations by the peak value 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.

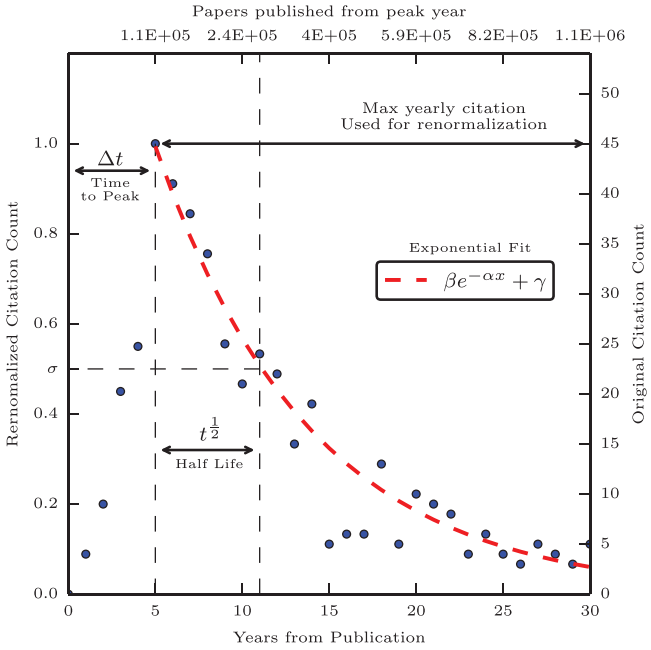

Fig. 2. Schematic representation of the citation evolution of a typical paper.

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

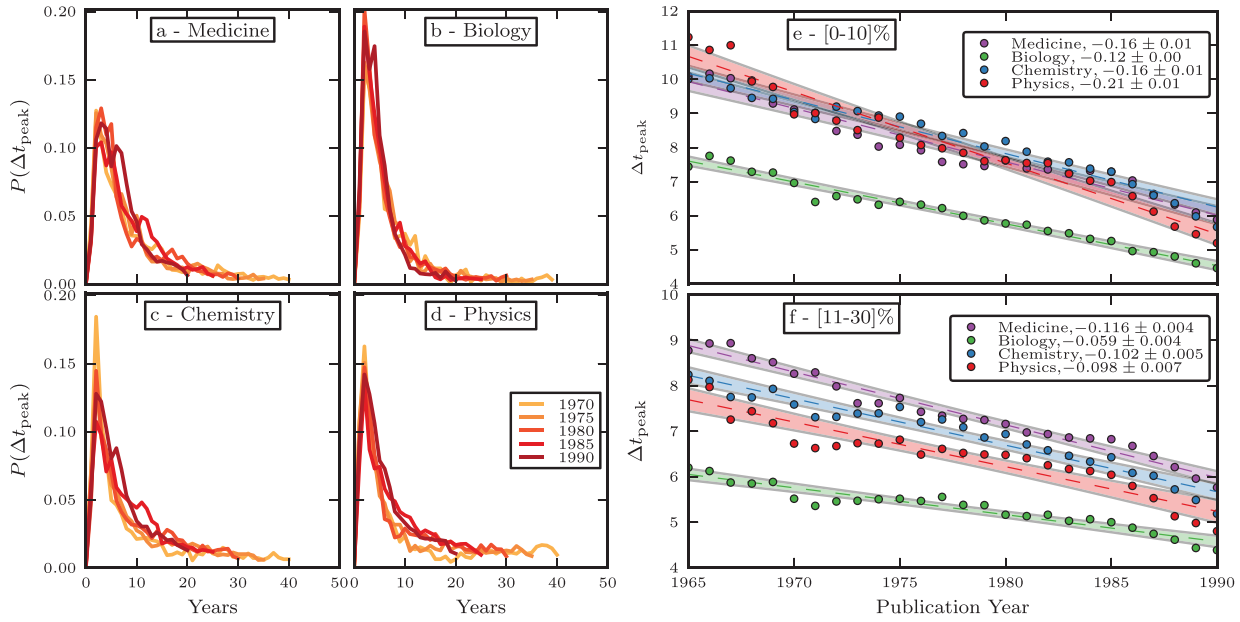

Fig. 3. Time to reach the peak attention Δt_{peak} is both field and time dependent. (a)–(d) Distribution of Δt_{peak} for papers in the top 10% published in the same year, for different fields and publication years. (e) and (f) Time evolution of the mean values of Δt_{peak} for top 10% and [11–30]% 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.

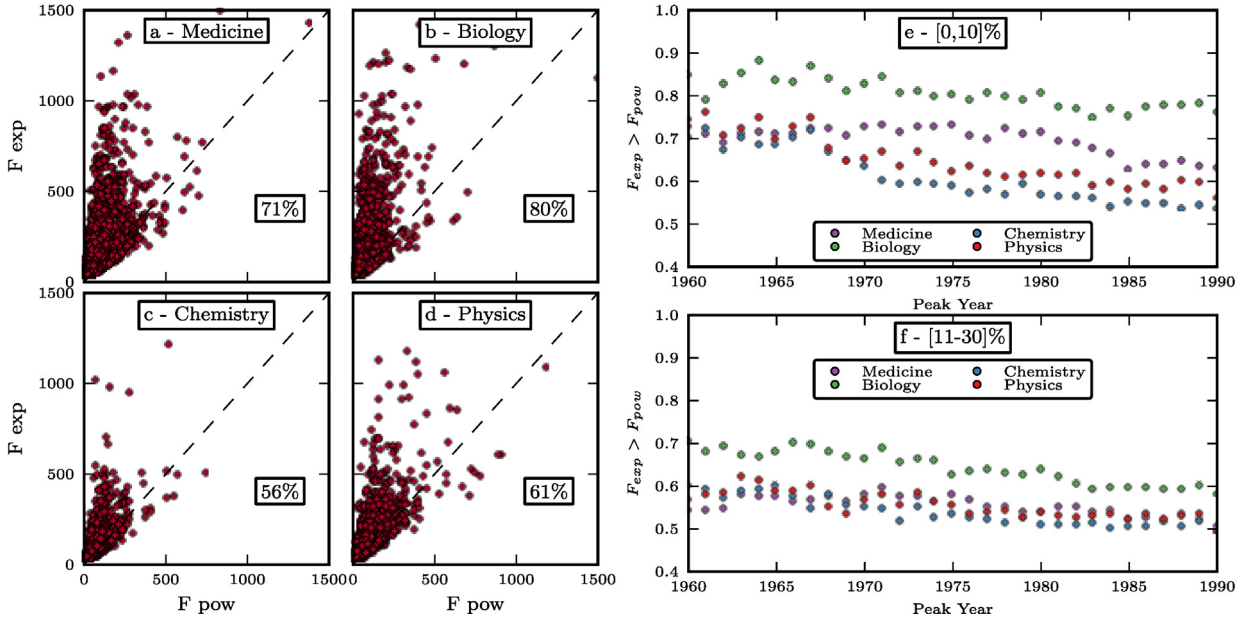

Fig. 4. Comparison of exponential fits with power law fits as described by the *F*-statistics. (a)–(d) Papers peaking in 1980, with the number in the box indicating the percentage of papers better fitted by exponentials than by power laws. In particular, it is worth noticing that there is a significant 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.

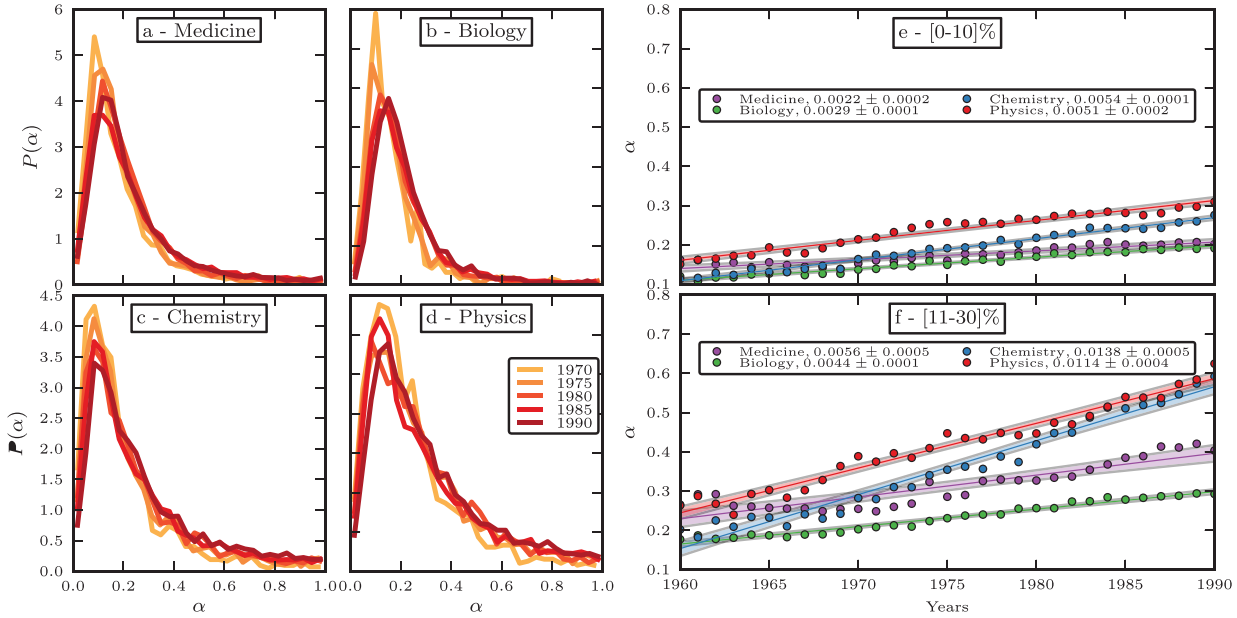

Fig. 5. Attention to publication is decaying faster in time. (a)–(d) Distribution of parameter α for exponential fits in different years for the four disciplines. For recent years the tail of the distribution becomes progressively fatter. (e) and (f) Time evolution of the median of the distributions 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

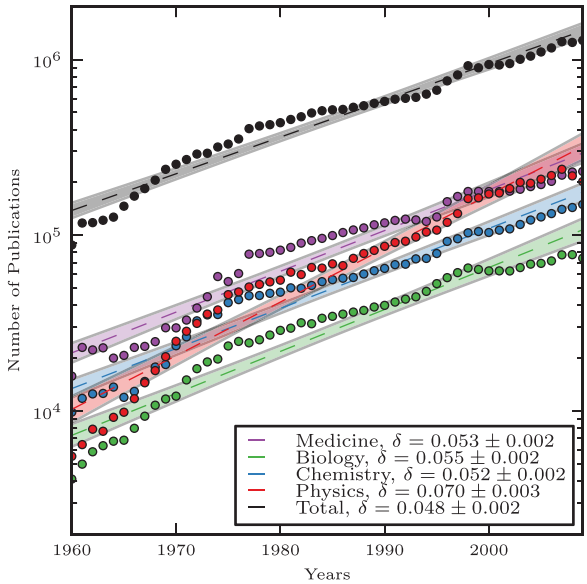

Fig. 6. Increase in the number of publications with time since 1960 along with exponential fits, 95% confidence intervals and rates.

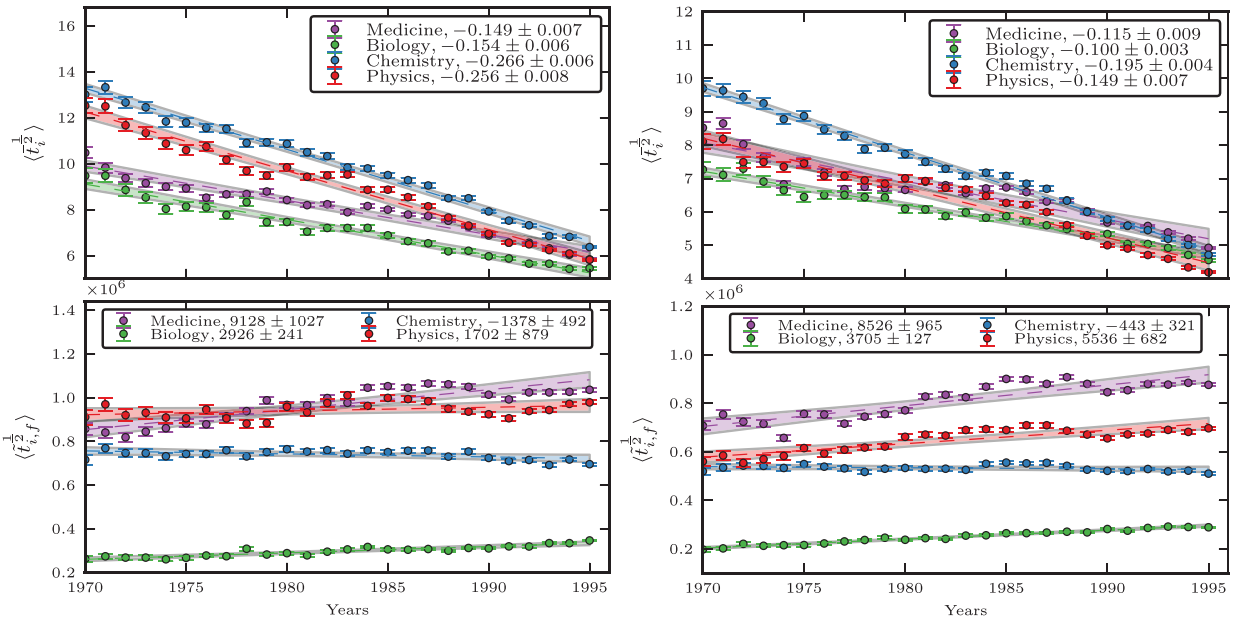

Fig. 7. The half-life of papers $t_i^{\frac{1}{2}}$ in terms of absolute time decreases linearly, whereas the rescaled half-life of papers $\widetilde{t}_i^{\frac{1}{2}}$ in terms of the number of publications is relatively constant. The panels show the evolution of $\langle t_i^{\frac{1}{2}} \rangle$ (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

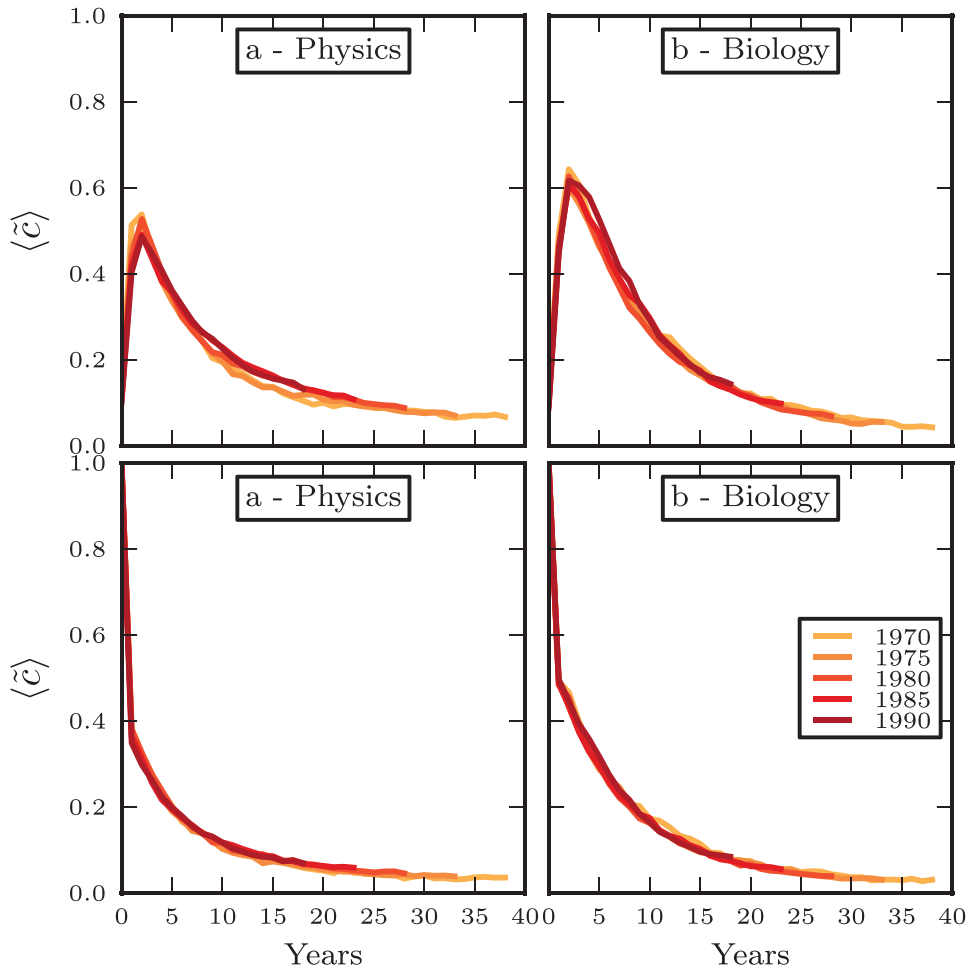

Fig. B.1. Averaged citation trajectories are calculated for papers in the [11–30]% window based on their total number of citations.

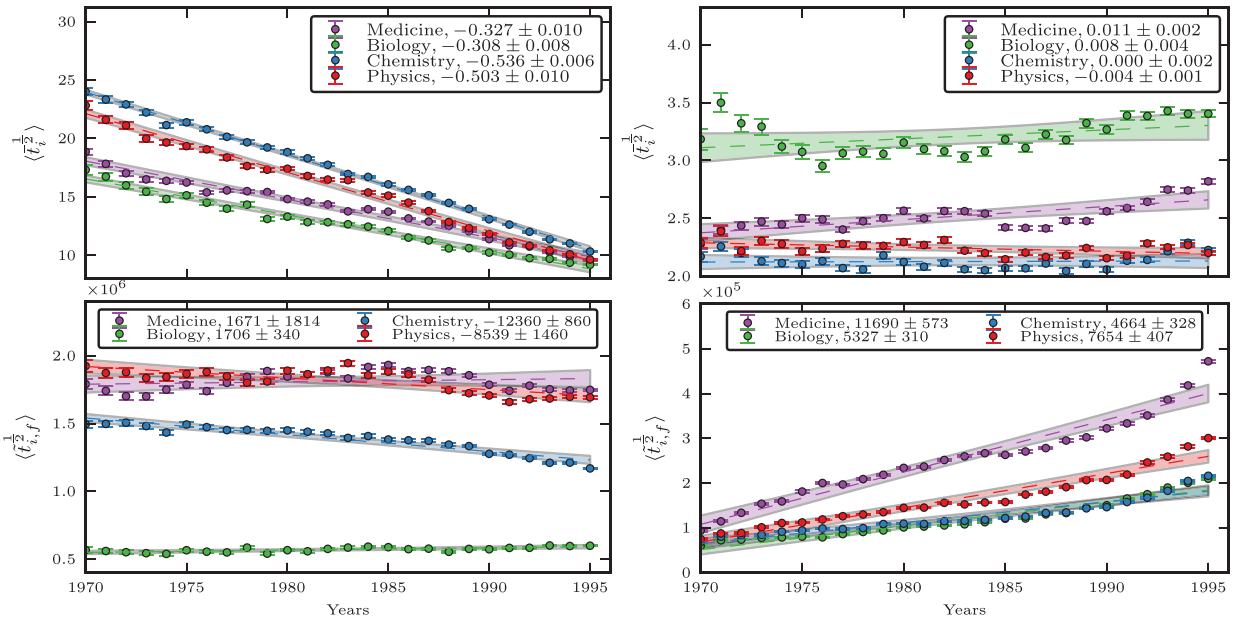

Fig. C.2. (Left) The half-life of papers $t_i^{\frac{1}{2}}$ with $\sigma = 0.3$. (Right) The alternative half-life of papers $\bar{t}_i$ with $\sigma = 0.5$.

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 the Journal of Informetrics, 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

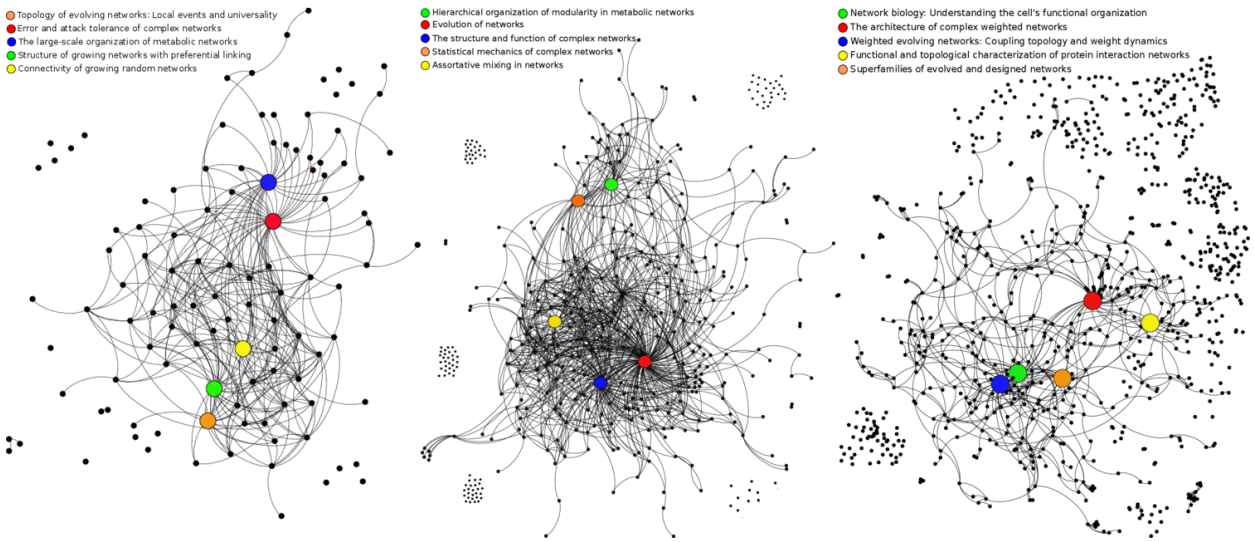

Figure 1: Ego-network for the paper *Emergence of Scaling in Random Networks* (Science 286, 509-512, 1999), by A.-L. Barabási and R. Albert. We consider windows of size $w = 2$ at t=1 (left), t=3 (center) and t=5 (right), where t is the number of years from publication. Therefore the windows are non-overlapping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

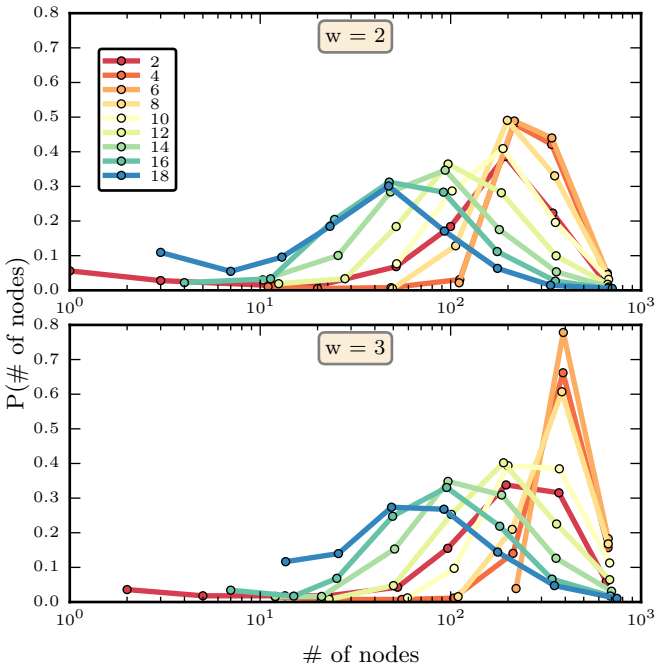

Figure 2: Distribution of EN size, with windows of 2 years (left) and 3 years (right). The different lines indicate different intervals t from the publication of the ego.

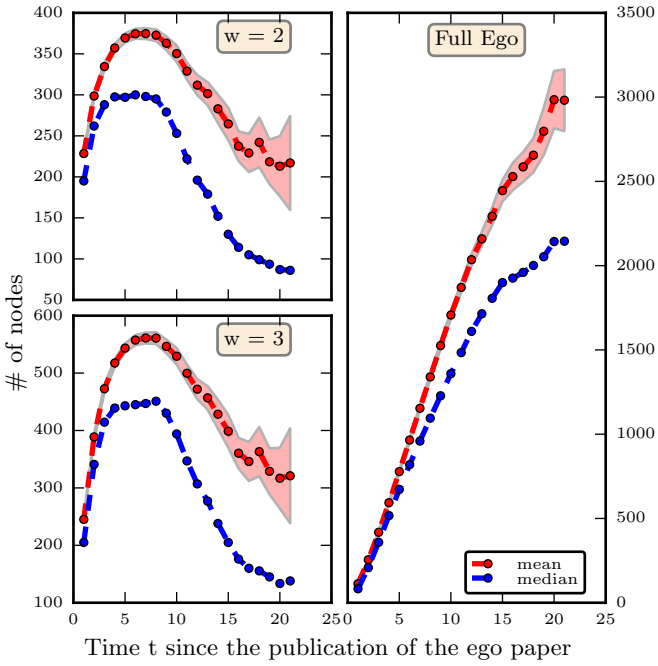

Figure 3: Evolution of mean and median absolute size of the ENs. We also show the standard deviation of the sampled mean. The panels correspond to $w = 2$ (top left), $w = 3$ (bottom left) and to the full EN (right).

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

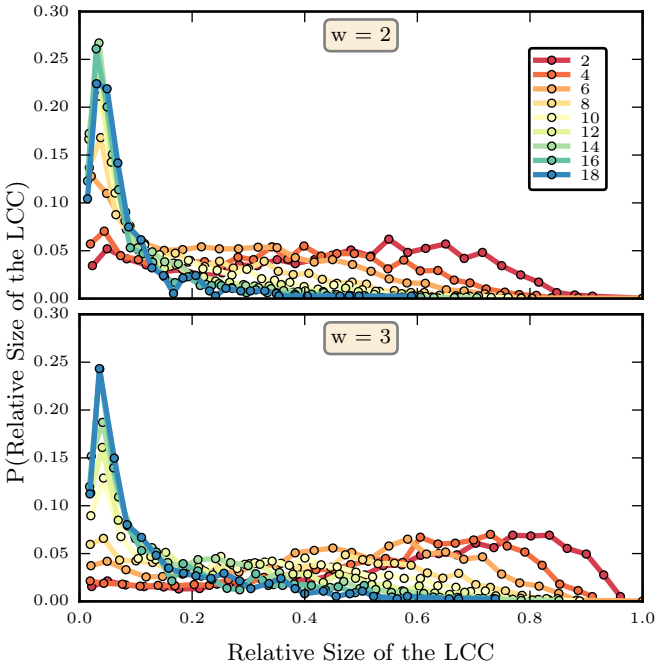

Figure 4: Distribution of the relative fraction of the largest connected component of the EN for all papers with windows of 2 years (top) and 3 years (bottom). The different lines indicate different distances t from publication.

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.

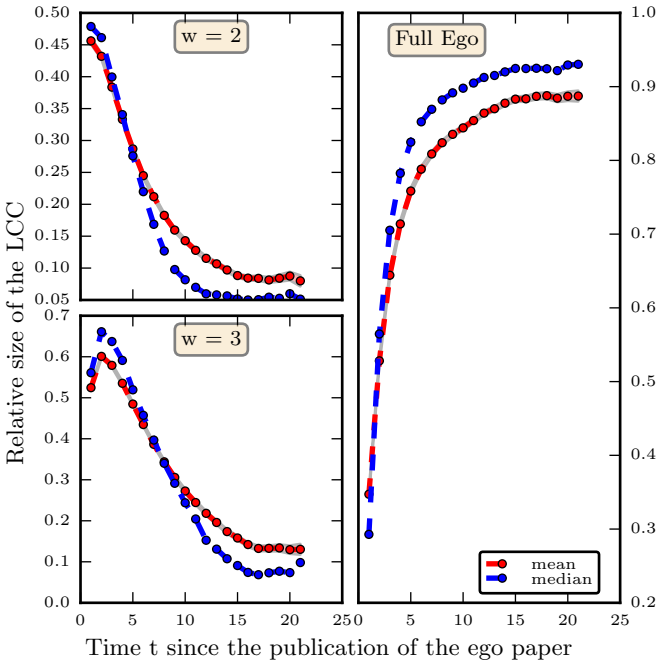

Figure 5: Time evolution of mean and median relative size of the LCC along with the standard deviation of the sampled mean for $w = 2$ (top left) and $w = 3$ (bottom left) and of the full EN (right).

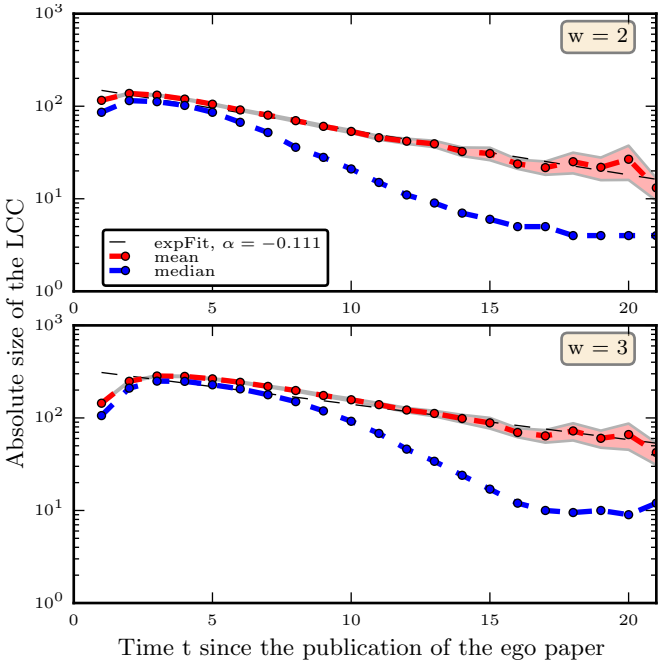

Figure 6: Time evolution of mean and median absolute size of the LCC for $w = 2$ (top) and $w = 3$ (bottom) along with the best fit to the exponential $t = \beta \exp(-\alpha t)$.

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.

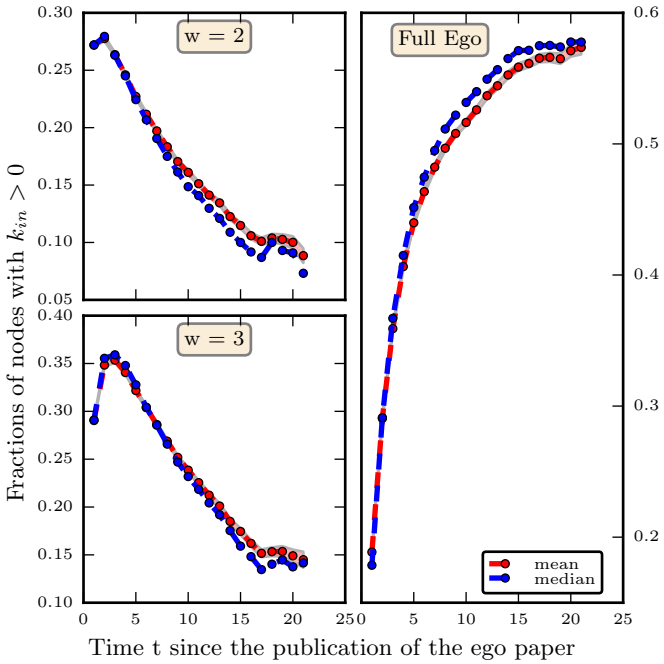

Figure 7: Time evolution of mean and median fraction of nodes with at least one incoming connection from the other nodes of the EN ($k_{in} > 0$) for $w = 2$ (top left), $w = 3$ (bottom left) and for the full EN (right).

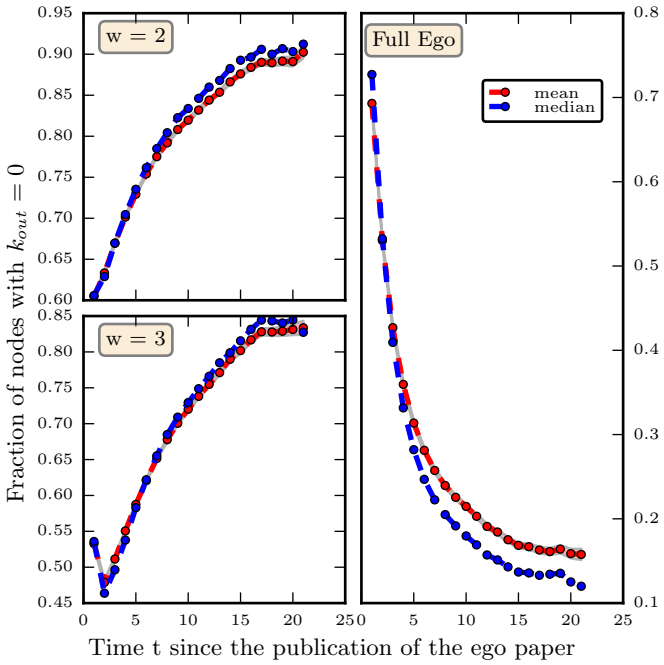

Figure 8: Time evolution of mean and median fraction of nodes without outgoing connections to the other nodes of the EN ($k_{out} = 0$) for $w = 2$ (top left), $w = 3$ (bottom left) and for the full EN (right).

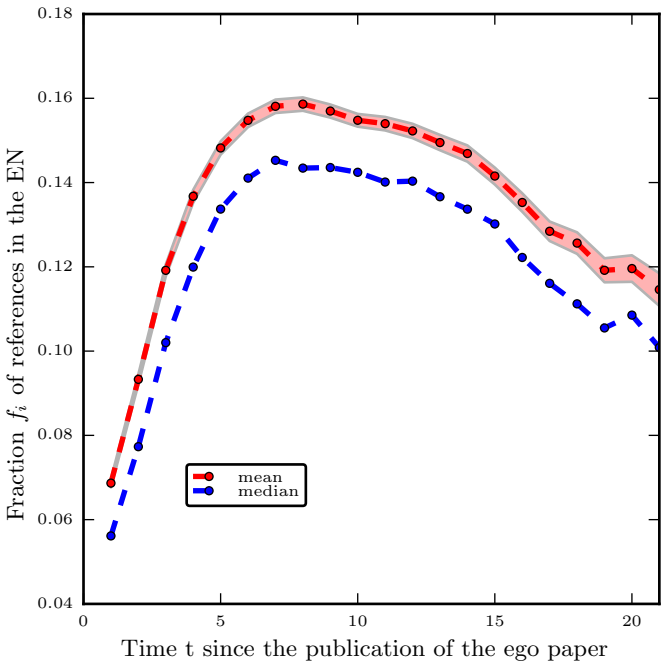

Figure 9: Time evolution of the mean and median of the fraction f_i of references of papers of the full EN belonging to the EN as a function of the number of years since publication.

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.

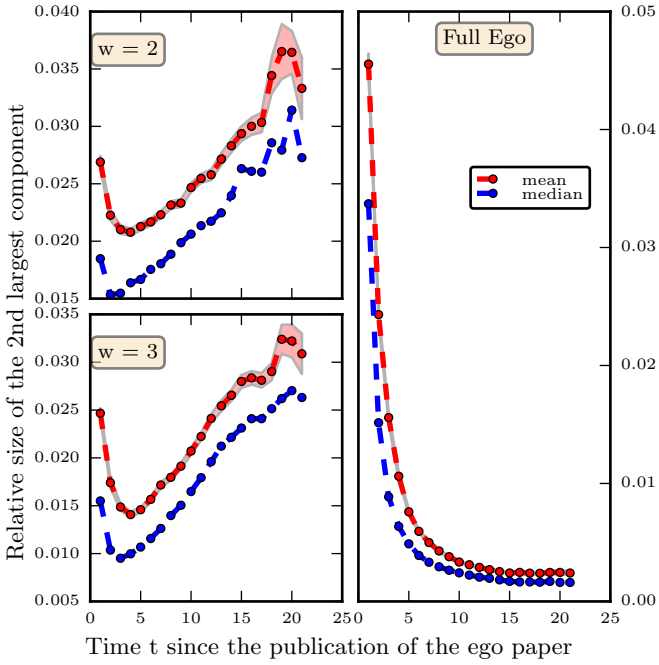

Figure A1: Time Evolution of mean and median relative size of the second largest component for $W = 2$ (top) and $W = 3$ (bottom) (left) and of the full EN (right).

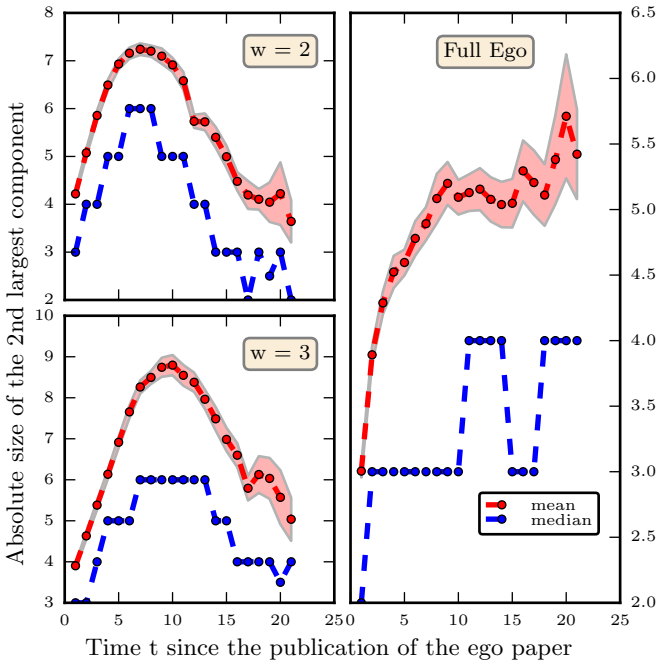

Figure A2: Time Evolution of mean and median absolute size of the second largest component for $W = 2$ (top) and $W = 3$ (bottom) (left) and of the full EN (right).

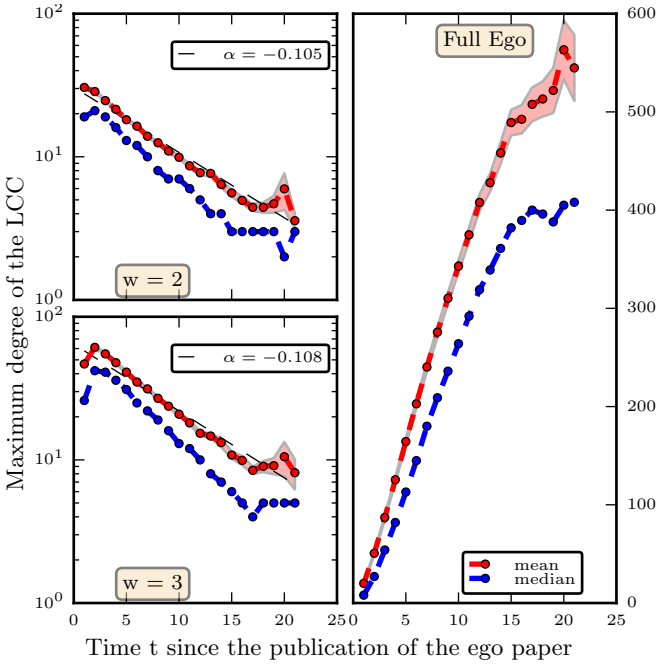

Figure A3: Time Evolution of mean and median relative size of the maximum degree of the lcc for $W = 2$ (top) and $W = 3$ (bottom) (left) along with an exponential fit and of the full EN (right).

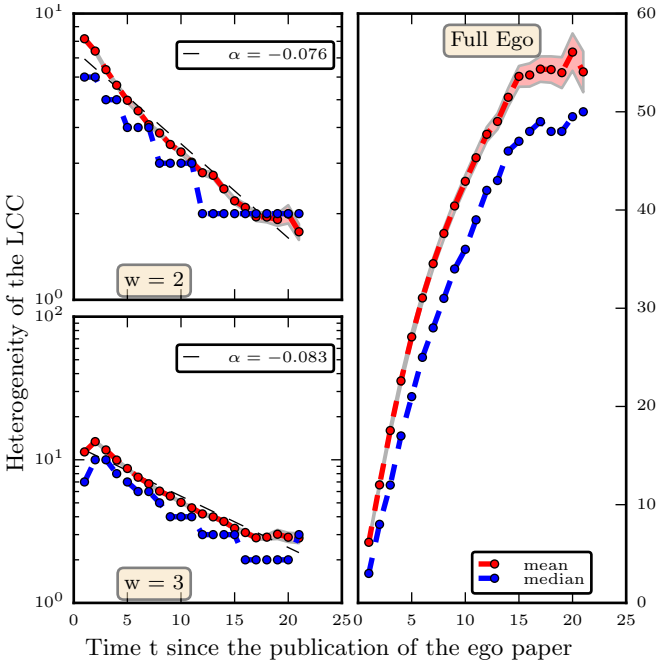

Figure A4: Time Evolution of mean and median relative size of the degree heterogeneity of the lcc, definied as $\frac{\sum d^2}{\sum d}$ for $W = 2$ (top) and $W = 3$ (bottom) (left) along with an exponential fit and of the full EN (right).

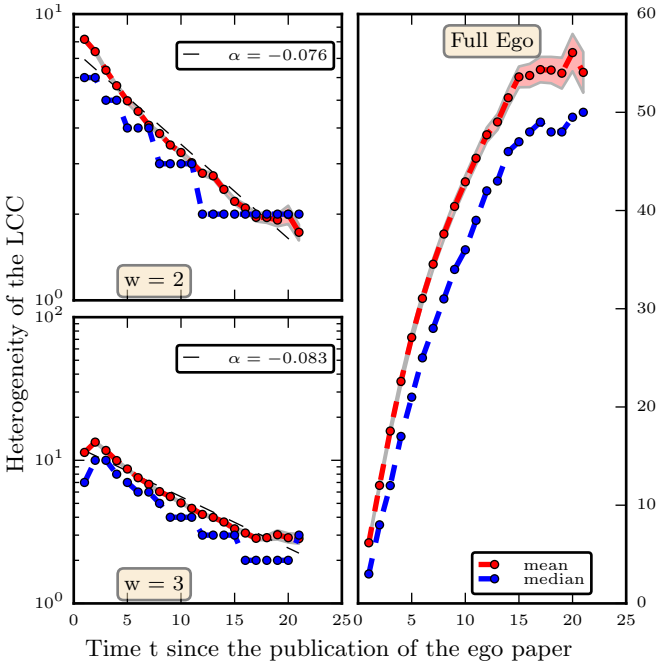

Figure A5: Time Evolution of mean and median relative size of the absolute size of the lcc for different citation volumes: $500 \leq c \leq 1000$ (left), $1000 \leq c \leq 2000$ (center) and $c > 2000$ (right). he higher the citation volume, the faster the decay. This seems to be caused by the fact that 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.

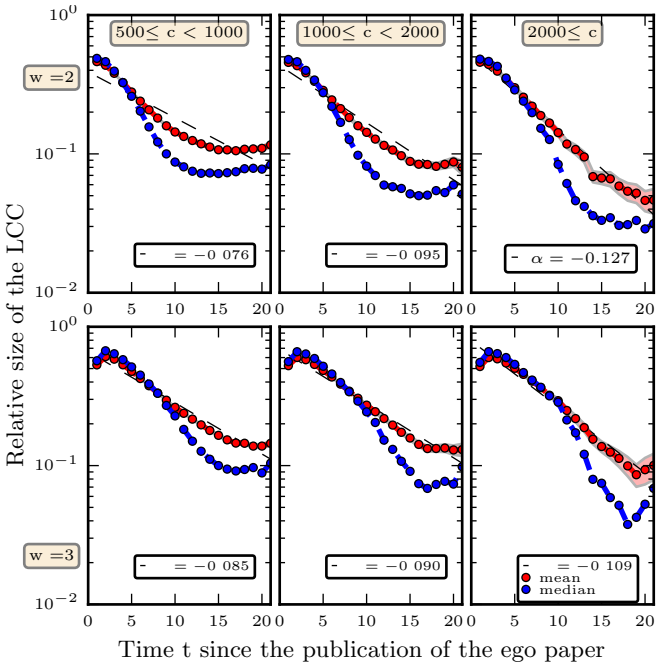

Figure A6: Time Evolution of mean and median relative size of the absolute size of the fraction of references that stay within the EN for different citation volumes: $500 \leq c \leq 1000$ (left), $1000 \leq c \leq 2000$ (center) and $c > 2000$ (right).

Publication IV

Pietro Della Briotta Parolo, Mikko Kivelä, Kimmo Kaski, Santo Fortunato. On the Shoulders of Giants: Tracking the Cumulative Information Spreading in Citation Networks. *preprint*, Submitted to Phys. Rev. X, June 2017.

On the Shoulders of Giants: Tracking the Cumulative Information 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: May 11, 2017)

A dominant paradigm in tracking the flow of information in science is to count direct citations between articles. However, scientific articles are built on articles they cite which in turn are based on cited articles. Similarly, the knowldge created in an article is not only retained by articles that direcly cite it. We investigate this cumulative process using two stylized models of information flow and a citation network of around 35 million publications. We find that ...

INTRODUCTION

Since the seminal paper from de Solla Price [1] quantitative analysis of scientific publications has been extensively used to study the structure and temporal evolution of science in general [2, 3] and in it specifically the changes of different scientific disciplines, fields, and subfields [4–6]. In particular, the citations between scientific publications are very interesting as they [encode various dependency relationships, information flow between articles, citations can have multiple meanings] [Citations have been used to map relationships not only between individual publications, but also between ...] Such citation networks describes linkages not only between individual publications, but also between publication forums or journals and between scientific disciplines or fields, which due to the long time-span of publication databases stretching over decades allows us deep insight to the evolution of science and its various disciplines and fields.

The main paradigm has been to look at direct citations. This thinking is exemplified by the quality measures that are based on direct citations, such as H index, impact factor and others.

In particular, many efforts have been produced in order to use citation networks in order to produce rankings of publications, starting with the famous H index [7]. Soon new, more complex measures followed, tackling the limits of the original measure, dealing with problems such as difference in publication fields [8], the quality of citing papers [9] as well as carreer length [10]. However, virtually all of these algorithms are based on the first layer of analysis, i.e. the citations received by the paper/author/journal whose rank is trying to be determined. This however is in contrast with the structure of science itself :science is a cumulative process in which one's results are intrinsically based on previous work. Therefore, when attempting to study its structure and behaviour it's necessary to look at the whole process and not just focus on a snapshot of the system. Science is an evolving social system, in which ideas will spread from scientist to scientist or group to group or institution to institution at different geographical locations thus allowing also to uncover geographical patterns of the idea spreading [11]. In addition one can make comparison between information spreading through the citation network and through social networks [12].

[Local viewpoint is mostly due to local data: in order to track flows further away you need to have global data. The sentence about databases could go here, as they help in this situation.] Some previous works have attempted to look at citation measures by using the global network either to rank publications [13] or individual scientists [14], finding correlation between local and global measures, yet with some very peculiar exceptions and with strong disparity.

In this work we want to continue along those lines, using a wider dataset and extending the analysis to journal, fields and subfields,focusing on studying the spreading of information, originating from a certain seed of papers, through the network. This is accomplished by selecting a starting group of papers coherent in terms of publication year and scientific field and diffusing the knowledge from the initial papers through the citation network, by following the citations that connect those original papers to their child nodes. This process is then repeated with the child nodes till the latest entries in our dataset (2008). By doing this we can show how the spreading of scientific information between different fields, subfields, journals, and individual papers takes place and how it changes or evolves in time.

This paper is organised such that we first describe the material and methods used in this study including the description of the data. Then we introduce two methdos for analyzing the spread of information, one more focused on the individual publications, which we call *Impact* and which is in nature similar to the pageRank algorithm. Then we look at a *Diffusion* process which gathers information from groups of publications. Our results show that the impact of indivudal papers (a global measure) is correlated to local properties, but showing an extreme variety. We also show how papers impact individual fields. For the diffusion process instead, we show that the speed at which information is being shared among fields is accelerating in absolute numbers, indicating a more rapid mixing among fields, yet showing a slowing down behaviour if we take in to account the increasing volume of publications.

MK: The diffusion/impact processes seem to be contained to a small subset of all papers even after long times. Is this because there are no paths to most pa-

pers or are the values leaking to large number paper just very small? That is, say that we start the process of diffusion/impact from a paper/subfield in 1970 and look at the distribution of values at 2008, is this distribution such that there are large differences in the values and that there is large number of papers involved, but only small minority have high enough values to have an effect?

I. MATERIAL AND METHODS

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 2010 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 milion publications in around 15 thousand scientific journals. We further map the subfields of the publications into major scientific fields []. MK: We can put the full tables as supplementary material. See how we should cite this. Also, we should cite any other articles using this field assignment.

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 $X\%$ of 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 2008. Previous literature [15] 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.

MK: Is the data until 2010 or 2013? Both years are mentioned. Check which one is correct.

MK: I modified the network construction part a lot. Check that everything is correct. Add the percentage X.

Impact 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 even quatify the amount of information 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 *pulling* process. In this starting from an original seed publication we attribute to it an initial value of Impact I of value 1, while all other publications have an initial value of 0. We then update the impact values of article published after the original one in chronological order such that node j pulls scientific value, if present, from articles it cites i and updates it own impact value as follows:

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

where k_j^{in} is the in-degree (or, number of references) of the article j.

In this pulling mechanism we consider the relative impact of the cited paper in the citing paper such that the citing paper inherits the impact of the cited papers but divides each pulled impact equally among the number of publications present in its reference list. A hypothetical publication with only one reference will pull the whole impact from the cited paper as its scientific results are entirely based on that one 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 pulled to the review.

MK: add: the amount of information/impact that is present in the articles at each time instant is not constant, but it can be copied as the results of an article is copied to many citing articles. It can also decrease in time when only small number of articles with big reference lists are citing it, or even completely disappear if nobody is citing an article.

A. Impact in the data

For each original seed publication we are thus able to see how much it has had an *impact* in the scientific world at different stages. Fig.1 shows an example of this by looking at the impact history of Roy J. Glauber's seminal paperon photons correlations [16] published in 1963, which eventually led to him winning a nobel prize, by comparing the results obtained by our method either by looking at the global or local scenario. Each column shows the impact of the paper in different fields/subfields/journals as well as the total impact, also showing the difference between the local and global scenario. In the left column we allow the initial impact to spread through the whole network, while in the right column we stop the process at the first layer of publications. We can see that while there is a certain similarity between the two columns, the right one is extremely changeable and relies heavily on the number of citations

received by the paper. In the global scenario instead, we see a continuum in the history of the impact of the paper, showing a gradual contribution of the paper to the field of "Mechanical" (orange in the field panel), which is not shown in the local measure. Similarly, we can see that the contribution to Engineering (light orange, third largest field) in the local scenario receives an initial spark and then occasional random sparks much later. In the global scenario instead the contribution to engineering continues gradually, yet firmly across the years. Looking at the subfield panel we can see the gradual growth of "Optics" (red color), which is responsible for the aforemtioned impact in Mechanical. In the local scenario, we see virtually no contribution instead, with occasional sparks in the fields arriving only after 10 years from publications. Similarly, the contribution to Optics is mainly due to two journals: Physical Review A (red) and Optics Community (dark yellow, 4th largest in total volume among journals), which once again are not present in the local scenario.

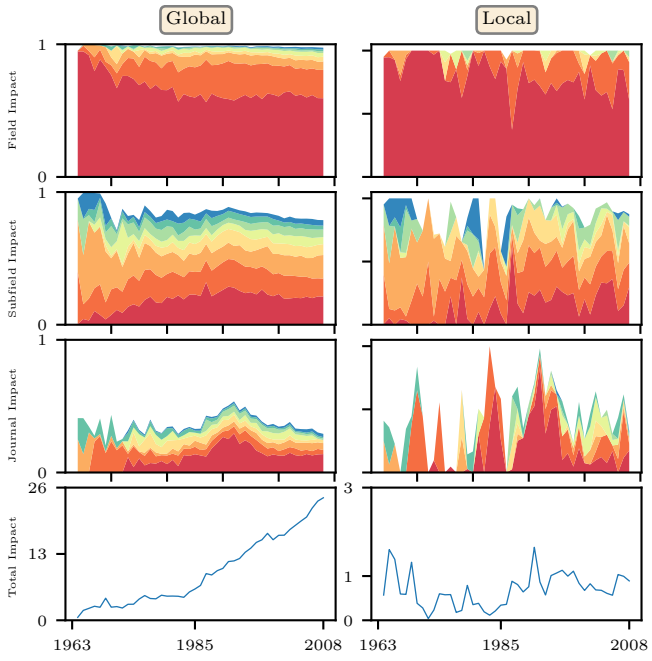

FIG. 1. Impact of Glauber's nobel paper 1963 to 2008. The left column shows impact measures linked to the global process, while the right column focuses only on the first layer of citations. The bottom panels show the total impact, measure as in equation 1. The top three panels show the relative distribution of impact among fields,subfields and journals (from top to bottom). The colors represent the same field/subfield/journal in both columns and are position in order (from bottom to top) of impact summed across years.

We calculated impact values for all papers in our datasets with at lest 20 citations, published between 1970 and 2008. This sums to a total of 6268678 publications. We also gathered the impact values for 74 "nobel" papers that we have selected, which are the seminal papers in nobel discoveries. For each publication we were thus able

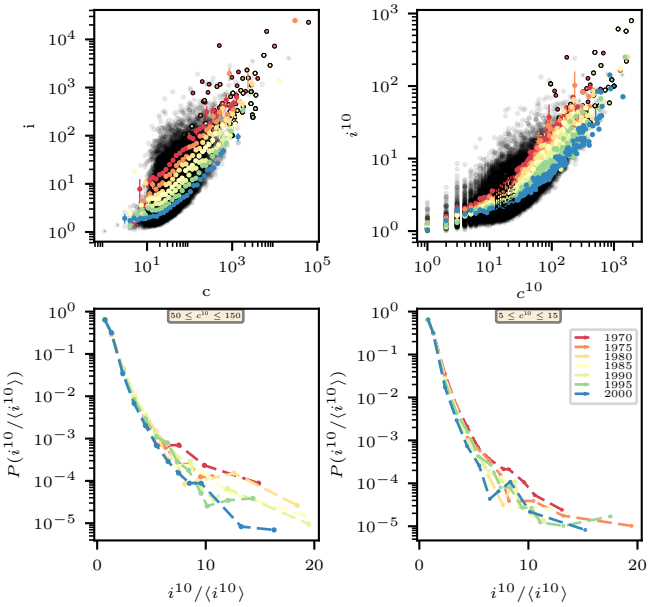

FIG. 2. Impact panel

to calculate the cumulative impact of the paper as well as the snapshot at each possible year, by summing only up to a desired year. This allows us to put ourselves in the point of view of any year in the past and to compare the impact of papers not only at the final step, but also along the history of the paper, thus removing the cumulative advantage of older papers.

Fig.2 shows a summary of the results. The top left figure shows the correlation between citations and total impact in different years, as well as the average binned data (full dots) and the nobel data (circled dots). This figure resembles the results in [13], with a power law correlation between the two values, but at the same time showing a huge variance of results within the same bins in terms of citations. Especially in the central range of citations (10 to 100) the distribution of impact can span numerous orders of magnitude. This indicates, as expected, that the number of citations per se is not sufficient to summarize fully the impact 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 nobel papers fall in the top right corner of the figure, indicating high values of both citations and impact. However, there are clearly a group of nobel publications that are well above the average of their year and citation bin, coherently with the fact that one of the main reasons behind a nobel discover is a significant contribution to the scientific world. The top right figure shows the same correlation but with

both values calculated after 10 years, in order to be able to compare the quantities at the same point in time. We can see that the same power law correlation exists, as well as a signficant higher impact for older publications within the same citation bin. Ths shows that also on a shorter time scale, publications with very similar number of citations, manage to have a extremely varied impact in the scientific world. Once again, nobel papers show to be overachievers, having impact 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 lenght of reference lists, which cause the denominator in Eq.1 to reduce the amount of impact in the citing papers. In fact, in the figure in the bottom left corner we can see the distribution of i^{10} values for papers with citations between 50 and 150, divided by the average of the bin. We can see that, across decades, the relative distrbution is exponential and stable, indicating that despite a change in average i^{10} values, the distribution relative to average remains constant in time. In the bottom right panel we see the same distributions but for less cited papers (between 5 and 15), showing the same pattern, yet with the collapse of curves being less robust. However, just like in the previous case, the deviations appear only in the tail of the distribution, indicating that they might be just be outliers.

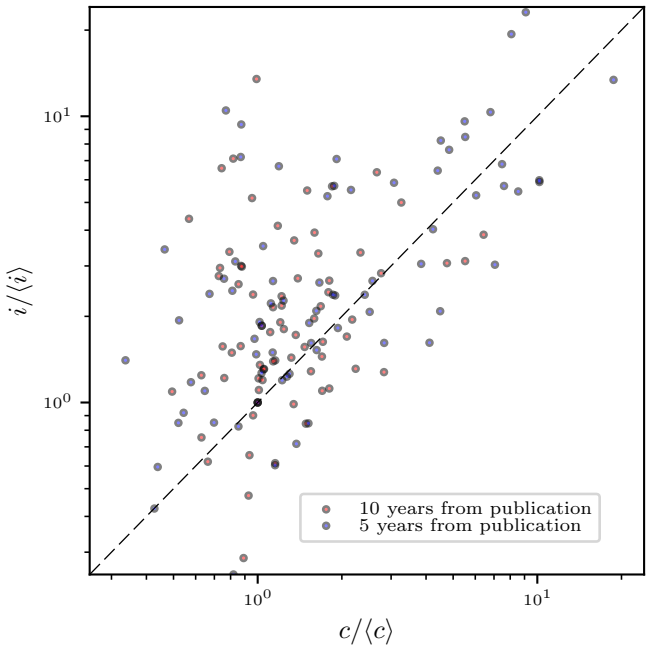

FIG. 3. Citation vs Impact outperformance of nobel papers, compared to similar papers in terms of citations after 5 and 10 years.

Fig. 3 see the correlation of the outperformance of nobel papers in both impact and citations. For each nobel paper we return the total impact 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 impact 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). 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.

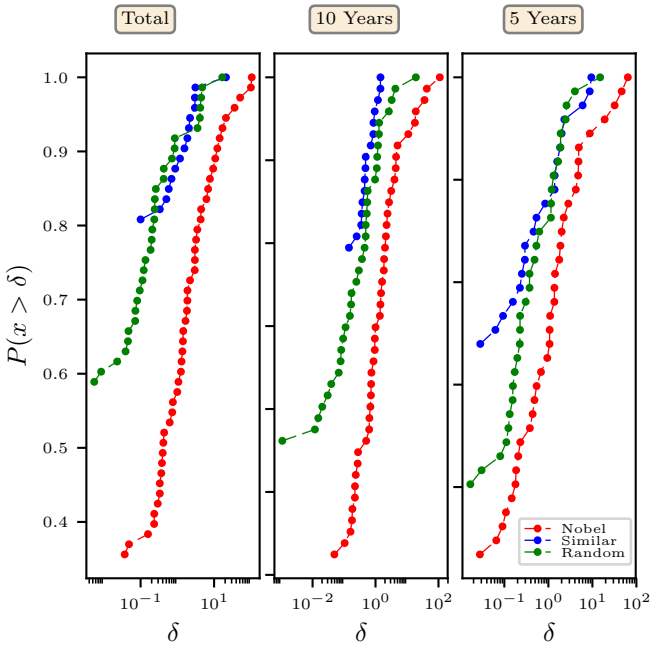

FIG. 4. Cumulative distribution of δ for nobel papers, paper within a 3% in citation volume in the same time interval, compared to nobel papers and for random papers. The insert instead shows the absolute number of position gained for papers with a positive gain.

We can also calculate rankings for papers published in the same year both in terms of citation and impact, where the paper with the lower ranking has the highest value of citations or impact within the selected group of papers. We define the relative change in ranking is as $\delta = \frac{cRank - iRank}{iRank}$. A negative δ indicates a loss of positions (i.e. lower impact ranking than citation ranking). The division by $iRank$ guarantees that papers with high impact ranking receive a higher δ compared to a similar paper who gains the same number of positions but lower in the rankings. Fig.4 shows the cumulative dis-

tribution 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 climb positions from their citation rankings. Furthermore, we can see from the figure that ca. two thirds of the Nobel papers are gaining positions, with the number being constant in time, while similar Papers have gone from 35% after 5 years to a minimum of 20% overall. 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. Such increasing difference in time, despite similar number of citations show how the impact measure is better able to grasp the cumulative contribution in science that the Nobel publications have had.

Table I shows the best performances in terms of change in δ for all papers in our dataset. The list offers a very intersting analysis of relatively minor papers who have, however, managed to contribute to the development of science by inspiring further research.

It is interesting to note how many of the papers in the first positions are from the 70s and are linked to the field of Gentics. 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 fith 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) disoveries in the mathetmatical 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. The list also shows evidence of the relatively new field of complex systems, 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.

δrank	c	I	cRank	IRank	Year	Title
1	31	2905.67	37588.5	5	1974	Hybridization On Filters With Competitor Dna In Liquid-Phase In A Standardand A Micro-Assay
2	51	2787.96	23366	5	1976	Nucleotide And Amino-Acid Sequences Of Gene-G Of Phix174
3	20	2075.07	62269	18	1975	Invitro Polyoma Dna-Synthesis - Inhibition By 1-Beta-D-Arabinofuranosyl Ctp
4	19	981.77	88381	32	1980	Inhomogeneous Superconducting Transitions In Granular A1
5	82	372.78	26353.5	10	1997	An Adjustment To The 1997 Estimate For New Prostate Cancer Cases
6	28	2766.22	28047	11	1970	Molecular Hybridization Between Rat Liver Deoxyribonucleic Acid And Complementary Ribonucleic Acid
7	34	770.16	63260	25	1985	A Novel Method For The Detection Of Polymorphic Restriction Sites By Cleavage Of Oligonucleotide Probes - Application To Sickle-Cell-Anemia
8	18	851.82	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	17	35.92	131750	72	2004	A New Method Of Predicting Us And State-Level Cancer Mortality Counts For The Current Calendar Year
10	22	418.88	114723	67	1988	Zn Related Electroluminescent Properties In Movpe Grown Gan
11	77	731.16	26020	16	1989	Structure And Intercalation Of Thin Benzene Derived Carbon-Fibers
12	109	1381.85	12231.5	8	1985	The Oxygen Defect Perovskite Bala4Cu5O13.4, A Metallic Conductor
13	45	2084.81	19801	13	1972	Translation Of Encephalomyocarditis Viral-Rna In Oocytes Of Xenopus-Laevis
14	136	3184.38	4216.5	3	1974	Amplified Ribosomal Dna From Xenopus-Laevis Has Heterogeneous Spacer Lengths
15	30	1401.83	42143.5	30	1975	Classification Of Acute Leukemias
16	37	76.43	62485.5	48	2002	Wild Topology, Hyperbolic Geometry And Fusion Algebra Of High Energy Particle Physics
17	26	898.19	58242.5	46	1978	Relation Between Mobility Edge Problem And An Isotropic Xy Model
18	49	598.51	42981.5	34	1986	Transcriptional And Posttranscriptional Roles Of Glucocorticoid In The Expression Of The Rat 25,000 Molecular-Weight Casein Gene
19	20	401.84	114240	91	1986	The Use Of Biotinylated Dna Probes For Detecting Single Copy Human Restriction-Fragment-Length-Polymorphisms Separated By Electrophoresis
20	29	380.5	92031	74	1989	A Solid-State Nmr-Study On Crystalline Forms Of Nylon-6
21	21	575.88	89271	74	1982	Multiple Opportunistic Infection In A Male-Homosexual In France
22	57	2804.23	11535	10	1970	Synthesis Of Ribosomal Rna In Different Organisms - Structure And Evolution Of Rrna Precursor
23	126	232.17	11227.5	10	1999	Small-World Networks: Evidence For A Crossover Picture
24	119	1676.79	12064.5	13	1987	Superconductivity At 52.5-K In The Lanthanum-Barium-Copper-Oxide System
25	32	414.08	71919	82	1986	Fractals - Wheres The Physics
26	67	600.64	28044	33	1986	The Complete Structure Of The Rat Thyroglobulin Gene
27	62	1141.37	21222	25	1979	Interference Detection Among Solids And Surfaces
28	19	623.99	81374	99	1979	Comparison Of The New Miniature Wright Peak Flow Meter With The Standard Wright Peak Flow Meter
29	84	2215.65	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	82	375.26	26908.5	34	1992	Materials Science - Strength In Disunity

TABLE I. Publications with highest δ.

Table II shows the ranking by citations only, which is dominated by Medicine, Biology and Genetics, thus failing to take into account significant areas of science which are not as highly cited as the aforementioned ones.

Diffusion

On top of the impact method, we also applied a *diffusion* method in our citation network. The idea in this case is to focus on groups of seed papers (ideally from the same journal/subfield/field) and to spread the scientific value of the original papers equally across all their child nodes. While the impact method i based on the paper level, this method is meant to look at grouping of papers.

We initialized a set of N seed papers to which we have assigned the same initial value $v_i = 1/N$. Alternatively one can assign initial values assimetrically, by looking at how many ciations each paper has received in the first 5 years (plus one, to take care of citationless papers), this way we have a proxy of how successful the paper has been in general (see our paper) and we avoid using information potentially outside of the choherence of the system by looking to deep in the future. In this scenario the initial value is defined as $v_i = \frac{c_i^5}{\sum_j c_j^5}$ As a start, 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.

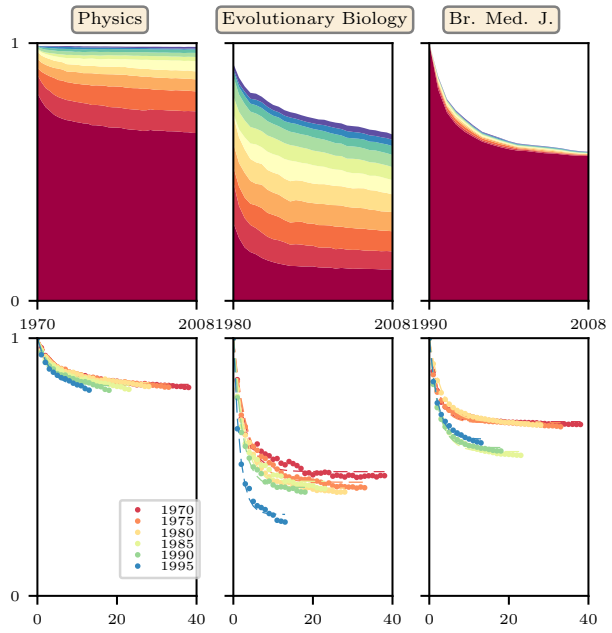

FIG. 5. Example of the diffusion method (top) and data fitting (bottom). The top panels show the diffusion of scientific value for Physics in 1970, Evolutionary Biology in 1980 and for the British Medical Journal in 1990.

Fig 5 (top) shows a tyipcal scenario for all types of initialization and pushing scenarios for the field of Physics, the subfield of Evolutionary Biology and for the British Medical Journal. 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 other fields/subfields/journals with the most value across time for each case. 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 tags a paper has). As time goes by the pushing method

pushes the scientific value 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 groups we have looked only at the amount of value retained by each field. For each start year we take the yearly values and we renormalzime them wit 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: $v(t) = (1-\beta)e^{-\alpha t} + \beta$. This allows to quantify numerically both the rate of change of the value in the initial years (through α) and the final plateau (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 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 $1 - \frac{1-\beta}{2}$:

$$\begin{aligned} 1 - \frac{1-\beta}{2} &= (1-\beta)e^{-\alpha t^{1/2}} + \beta \\ \frac{1}{2}(1-\beta) &= (1-\beta)e^{-\alpha t^{1/2}} \end{aligned} \tag{2}$$

allowing us to use the classical definition of half life as :

$$t^{1/2} = \ln(2)/\alpha \tag{3}$$

Overall, we can see that the same general pattern is retained: the original scientific 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 loss of scientific value 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 multidiscplinary has the lowest half life in both years, coherently with the fact that it is meant to be a field open to share 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, also coherently with the fact that it has become recently a structure field of its own. Also, it is interesting to notice some qualitative patterns between so called "hard" and "soft" sciences. While in 1970 we can see some humanistic fields showing very high values for their half lives (Phylosophy,History, Anthropology, Literature, 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. 6 (left column) shows the evolution in half life for the some of the major fields, all subfields and a filtered list of journals. The top panel shows the relative change in half life, compared to 1970. We can see that all fields show a cosntant decreasing pattern, losing between 20 and 60 percent of their value. The two panels for subfields and journals instead show the cumulative distribution of the half life values. As we can see, for both cases, the more recent distributions (green tones) are above the older ones (red tones), showing that the values have decreasing.

However, previous studies show [17] that years may not be the best choice to measure the rate at which changes happen in science, in favor of using 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 introduced in the system. Furthermore, while the publication growth is exponential, its growh rate is sufficiently small: $N(t) \approx N_0 \exp \delta t$ with $\delta \sim 0.05$ across all fields, allowing us to keep the same functional forms for the fits:

$$\begin{aligned} v(N) &= v(\int_{t_0}^{t_N} N(t)dt) \\ &= (1-\beta)e^{-\alpha N_0 \int_{t_0}^{t_N} \exp^{\delta t}\, dt} + \beta \\ &\approx (1-\beta)e^{-\alpha N_0 \left[\frac{1+\delta t}{\delta}\right]_{T_0}^{T_N}} + \beta \\ &\approx (1-\beta)e^{-\alpha \delta^{-1}(N_0(1+\delta(T_N - T_0))} + \beta \\ &\approx (1-\beta)Ke^{\alpha N_0 \Delta(T)} + \beta \\ &= (1-\beta)Ke^{\alpha^* \Delta(T)} + \beta \end{aligned} \tag{4}$$

Therefore now we are able to quantify the half life not in terms of years, but rather in terms of number of papers published in the meantime.

Fig.6 (right column) shows the same results for the renormalized scenario, with time being replaced by the renormalization method mentioned in Eq.4. Interestingly, the decreasing behaviour is no more dominating, with only one field showing (Chemistry) showing a downard pattern. All other major fields intstead, either remain constant or show a significant increase 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.

CONCLUSIONS

	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. 2. Pushing along with the relative change for each value. MK: pushing along with relative change for each value?

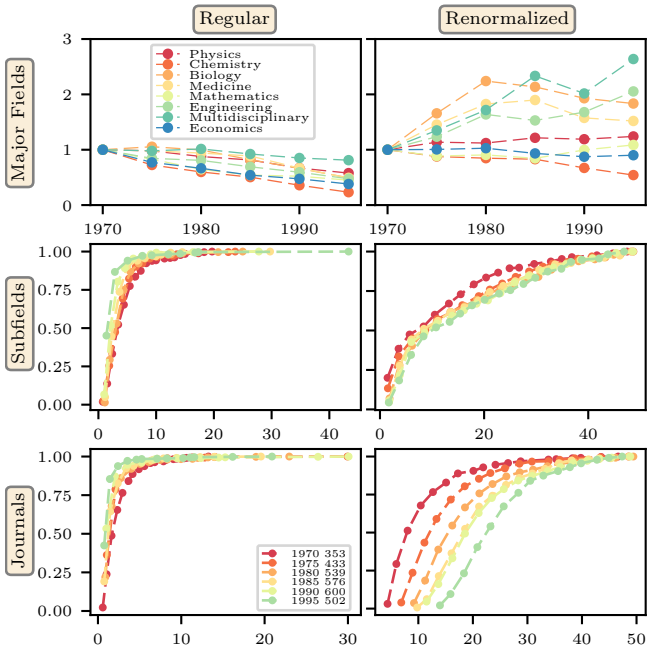

FIG. 6. Changes in half life in time for the regular (left column) and renormalized scneario (right) and for different grouping of papers.

MK: Future direction: look into past similar to impact, for a given article find out where the influence is coming.

REFERENCES

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

[2] H. Small, Journal of the American Society for Information Science **24**, 265 (1973).

[3] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, Internet Mathematics **6**, 29 (2009), http://dx.doi.org/10.1080/15427951.2009.10129177.

[4] M. E. J. Newman, Proceedings of the National Academy of Sciences **98**, 404 (2001), http://www.pnas.org/content/98/2/404.full.pdf.

[5] 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.

[6] M. Herrera, D. C. Roberts, and N. Gulbahce, PLOS ONE **5**, 1 (2010).

[7] J. E. Hirsch, Proceedings of the National Academy of Sciences of the United States of America **102**, 16569 (2005).

[8] 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.

[9] M. Bras-Amors, J. Domingo-Ferrer, and V. Torra, Journal of Informetrics **5**, 248 (2011).

[10] H. JE, (2005), doi:10.1073/pnas.0507655102.

[11] R. K. Pan, K. Kaski, and S. Fortunato, arXiv preprint arXiv:1209.0781 (2012).

[12] T. Kuhn, M. c. v. Perc, and D. Helbing, Phys. Rev. X **4**, 041036 (2014).

[13] P. Chen, H. Xie, S. Maslov, and S. Redner, Journal of Informetrics **1**, 8 (2007).

[14] F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani, Phys. Rev. E **80**, 056103 (2009).

[15] V. Larivire, . Archambault, and Y. Gingras, Journal of the American Society for Information Science and Technology **59**, 288 (2008).

[16] R. J. Glauber, Phys. Rev. Lett. **10**, 84 (1963).

[17] P. D. B. Parolo, R. K. Pan, R. Ghosh, B. A. Huberman, K. Kaski, and S. Fortunato, Journal of Informetrics **9**, 734 (2015).

Appendix A: Computational considerations

Appendix B: Description of the categories

Appendix C: Highest Cited Papers

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;
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;

TABLE I. Aggregation of TR subject categories in broader fields.

δrank	c	I	cRank	IRank	Year	Title
1	188969	57019.79	1	1	1970	Cleavage Of Structural Proteins During Assembly Of Head Of Bacteriophage-T4
2	8035	971.98	1	63	1971	The Assessment And Analysis Of Handedness: The Edinburgh Inventory
3	22793	11540.45	1	1	1972	Regression Models And Life-Tables
4	8693	1924.38	1	19	1973	Relationship Between Inhibition Constant (K1) And Concentration Of Inhibitor Which Causes 50 Per Cent Inhibition (I50) Of An Enzymatic-Reaction
5	8673	10627.8	1	1	1974	Film Detection Method For Tritium-Labeled Proteins And Nucleic-Acids In Polyacrylamide Gels
6	29897	24606.62	1	1	1975	Detection Of Specific Sequences Among Dna Fragments Separated By Gel-Electrophoresis
7	121947	19000.29	1	1	1976	Rapid And Sensitive Method For Quantitation Of Microgram Quantities Of Protein Utilizing Principle Of Protein-Dye Binding
8	62322	22686.34	1	1	1977	Dna Sequencing With Chain-Terminating Inhibitors
9	7179	1566.13	1	11	1978	Rapid Chromatographic Technique For Preparative Separations With Moderate Resolution
10	48152	10293.81	1	1	1979	Electrophoretic Transfer Of Proteins From Polyacrylamide Gels To Nitrocellulose Sheets - Procedure And Some Applications
11	9406	1764.12	1	6	1980	Ligand - A Versatile Computerized Approach For Characterization Of Ligand-Binding Systems
12	15202	3494.53	1	1	1981	Improved Patch-Clamp Techniques For High-Resolution Current Recording Fromcells And Cell-Free Membrane Patches
13	13539	4127.66	1	1	1982	A Simple Method For Displaying The Hydropathic Character Of A Protein
14	20817	7625.15	1	1	1983	A Technique For Radiolabeling Dna Restriction Endonuclease Fragments To High Specific Activity
15	13902	3335.6	1	2	1984	A Comprehensive Set Of Sequence-Analysis Programs For The Vax
16	16886	2143.56	1	4	1985	A New Generation Of Ca-2+ Indicators With Greatly Improved Fluorescence Properties
17	13581	2240.62	1	3	1986	Statistical Methods For Assessing Agreement Between Two Methods Of Clinical Measurement
18	56038	7085.36	1	1	1987	Single-Step Method Of Rna Isolation By Acid Guanidinium Thiocyanate Phenolchloroform Extraction
19	22123	1260.75	1	8	1988	Development Of The Colle-Salvetti Correlation-Energy Formula Into A Functional Of The Electron-Density
20	8191	552.05	1	32	1989	Gaussian-Basis Sets For Use In Correlated Molecular Calculations .1. The Atoms Boron Through Neon And Hydrogen
21	24360	3150.28	1	2	1990	Basic Local Alignment Search Tool
22	12799	1481.09	1	2	1991	Molscript - A Program To Produce Both Detailed And Schematic Plots Of Protein Structures
23	7970	1134.87	1	5	1992	The Mos 36-Item Short-Form Health Survey (Sf-36) .1. Conceptual-Framework And Item Selection
24	22380	1266.28	1	1	1993	Density-Functional Thermochemistry .3. The Role Of Exact Exchange
25	25752	1921.11	1	1	1994	Clustal-W - Improving The Sensitivity Of Progressive Multiple Sequence Alignment Through Sequence Weighting, Position-Specific Gap Penalties And Weight Matrix Choice
26	5293	571.37	1	4	1995	Genepop (Version-1.2) - Population-Genetics Software For Exact Tests And Ecumenicism
27	9146	616.3	1	2	1996	Generalized Gradient Approximation Made Simple
28	21200	1717.6	1	1	1997	Gapped Blast & Psi-Blast: A New Generation Of Protein Database Search Programs
29	11093	639.93	1	2	1998	Crystallography And Nmr System: A New Software Suite For Macromolecular Structure Determination
30	7357	440.28	1	2	1999	Mechanisms Of Disease - Atherosclerosis - An Inflammatory Disease

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