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

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

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

(Dated: June 8, 2017)

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

I. INTRODUCTION

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

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

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

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

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

II. DATA DESCRIPTION

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

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

III. PERSISTENT INFLUENCE PROCESS

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

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

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

where k_j^{in} is the in-degree (or, number of references) of the article j, and N_j is the set of out-neighbors. The normalisation guarantees that the sum of influence that the cited articles have on article j is constant and that the influence value never exceeds 1. When the process continues, the influence values dilute but at the same time it is spread to increasing number of articles.

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

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

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

A. Results

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

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

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

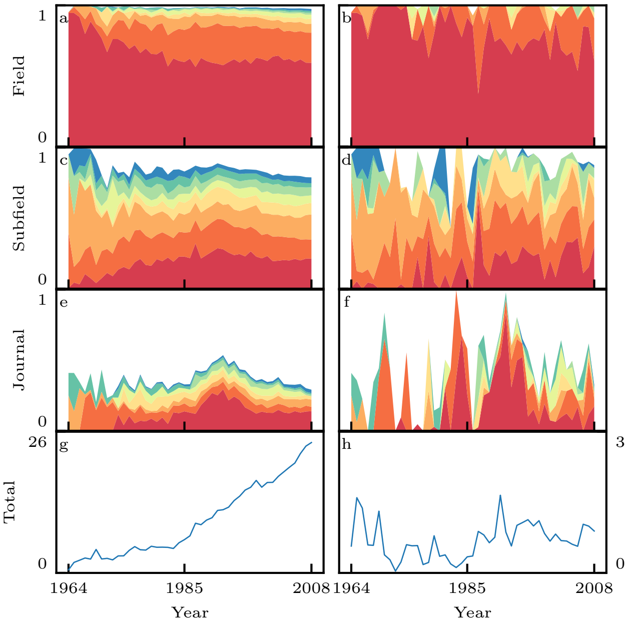

FIG. 1. Global and local influence profiles of Glauber's Nobel Prize winning paper from 1963 to 2008. The left column (panels a, c, e, g) shows the influence measures linked to the global process, while the right column (panels b,d, f, h) shows the influence calculated only from direct citations. Panels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

We calculated the influence values for all papers in our

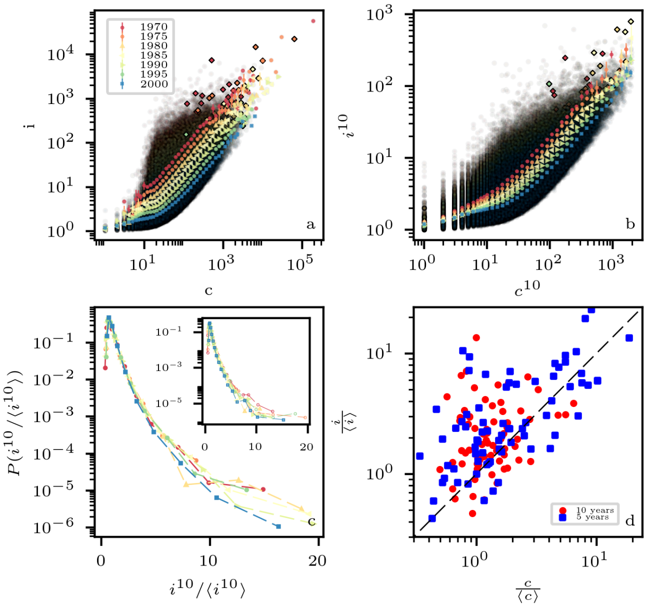

FIG. 2. Panel a shows the scatter plot between citation values and influence values for different years along with bin averages (filled markers). The diamond shapes represent Nobel papers, while the color is chosen according to the rounded closest value among the available years. Panel b shows the same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

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

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

The older paper have had more time to gather total influence in our data set, and to remove this advantage we calculated the cumulative influence value and citation count after 10 years from the publication date of each article (Fig.2b). A similar relationship as described earlier between the influence and citation count exists also in this case. Even the variance between influence values of articles with similar numbers of citations remains high (Fig.2c and insert). This shows that also on a shorter time scale, publications with very similar number of citations, manage to havedifferen a extremely varied impact in the scientific world. Once again, papers related to Nobel Prizes show to be overachievers, having influence also in this time scale significantly higher than the average for their number of citations. The difference shown in the average value between papers in different publications year is less strong than before and could be caused by an increase in the length of reference lists, which cause the denominator in Eq.1 to reduce the amount of impact in the citing papers. In fact, when comparing the influence distributions of articles with similar number of citations across years (Fig.2c) we can see that, across decades, the relative distribution is super-exponential and relatively stable, indicating that despite a change in average total influence values after 10 years (i^{10}), the distribution relative to average remains constant in time. Fig. 2 (d) shows the correlation of the outperformance of Nobel papers in both influence and citations. For each Nobel paper we return the total influence and citation values of papers within 10% of their citation count after 10 or 5 years and published in the same year. We then proceed to calculate the outperformance value of the Nobel paper by calculating the ratio between its own total influence and citation value and the average of each distribution for papers in the same citation bin. The figure shows the correlation between the outperformance in impact and in citations. In 73% of the cases (54 papers out of 74) the citation outperformance has been greater than the impact one. Also, the average impact outperformance after 5 years is 41% greater (4 versus 2.8), while after 10 years it rises to 66% greater (2.44 versus 1.47). Finally, while there are cases in which the outperformance has been greater in terms of citations, there are no extreme values in bottom right corner (high citation performance, low influence performance), while there is a solid vertical band in the opposite side of the plot, indicating a group of papers with low citation performance and high influence instead. This shows that Nobel papers, on average, outpeform papers who had similar citation counts after the same amount of time. This is due to the continued ability of the original Nobel paper (as well of its scientific offspring) to continuously propagate their ideas in the scientific community, as one would expect from such important publications. This result supports the finding of [27], where groundbreaking papers by important Nobel laureates were found to have a more "compact" network of child nodes over multiple layers.

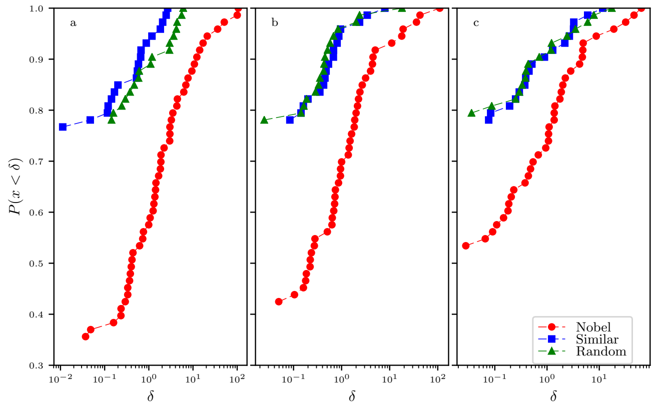

FIG. 3. Cumulative distribution of relative difference in persistent influence rank and citation count rank δ (given in Eq. 2). Three categories of papers are considered: Nobel papers (red circles), Nobel control set with papers within a 3% in citation volume in the same time interval (blue squares),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

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

$$\delta = \frac{R_c - R_I}{R_I}\,, \tag{2}$$

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

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

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

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

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

IV. DIFFUSION

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

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

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

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

A. Results

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

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

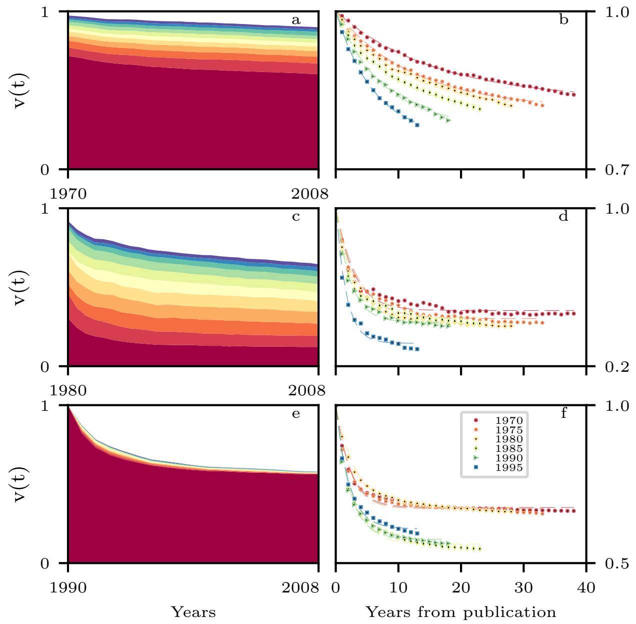

FIG. 4. Example of the diffusion method (a-c-e) and data fitting (b-d-f). The diffusion of scientific value for (a, b) Economics in 1970, (c, d) Evolutionary Biology in 1980, and (e, f) the British Medical Journal in 1990. The darkest tone of red shows always the amount of scientific value retained by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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

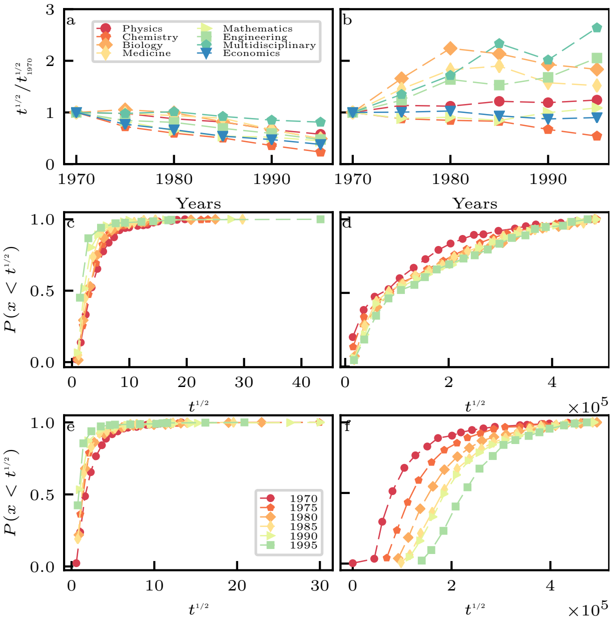

FIG. 5. Changes in half life in time for the regular (left column, panels a-c-e) and renormalized scenario (right, panels b-d-f) and for different grouping of papers. Panels a and b show the evolution of the half life for 8 different fields in units of the half life they had in 1970 in order to be able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

V. DISCUSSION AND CONCLUSIONS

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

We have introduced two simple methods to analyse how the knowledge created in an article, or in a group of articles, might percolate trough the scientific literature. These methods follow the tradition of modelling dynamics on networks, where a real observed network is used as a substrate where a stylized models progress is tracked. For example, this approach has been extensively used in network science to study epidemic spreading [43] and social dynamics [44]. In all of these cases the models are not expected to exactly mimic the real behavior, but the goal is to produce behavior that is accurate in the large scale. Our goal in this work was to introduce this approach to knowledge spreading in citation networks.

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

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

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

VI. ACKNOWLEDGMENTS

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

Appendix A: Computational considerations

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

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

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

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

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

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

Appendix B: Highest Cited Papers

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

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

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

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

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

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

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

tional Academy of Sciences **98**, 404 (2001), http://www.pnas.org/content/98/2/404.full.pdf.

[6] A. Barabsi, H. Jeong, Z. Nda, E. Ravasz, A. Schubert, and T. Vicsek, Physica A: Statistical Mechanics and its Applications **311**, 590 (2002).

[7] R. K. Pan, K. Kaski, and S. Fortunato, Scientific Reports **2** (2012), 10.1038/srep00902.

[8] F. Havemann, M. Heinz, and H. Kretschmer, Journal of Biomedical Discovery and Collaboration **1**, 6 (2006).

[9] B. F. Jones, S. Wuchty, and B. Uzzi, Science **322**, 1259 (2008).

[10] R. Sinatra, P. Deville, M. Szell, D. Wang, and A.-L. Barabási, Nature Physics **11**, 791 (2015).

[11] M. Rosvall and C. T. Bergstrom, Proceedings of the National Academy of Sciences **105**, 1118 (2008), http://www.pnas.org/content/105/4/1118.full.pdf.

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

[13] C. Hurt, Information Processing & Management **23**, 1 (1987).

[14] L. Bornmann and H.-D. Daniel, Journal of Documentation **64**, 45 (2008).

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

[16] E. Garfield, Canadian Medical Association Journal **161**, 979 (1999), 10551195.

[17] O. Penner, R. K. Pan, A. M. Petersen, K. Kaski, and S. Fortunato, Scientific Reports **3** (2013), 10.1038/srep03052.

[18] R. Adler, J. Ewing, and P. Taylor, Statistical Science **24**, 1 (2009).

[19] S. Alonso, F. Cabrerizo, E. Herrera-Viedma, and F. Herrera, Journal of Informetrics **3**, 273 (2009).

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

[21] P. D. Batista, M. G. Campiteli, and O. Kinouchi, Scientometrics **68**, 179 (2006), http://www.akademiai.com/doi/pdf/10.1007/s11192-006-0090-4.

[22] T. Braun, W. Glänzel, and A. Schubert, Scientometrics **69**, 169 (2006).

[23] L. Egghe, Scientometrics **69**, 131 (2006).

[24] R. K. Merton, American Sociological Review **22**, 635 (1957).

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

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

[27] D. F. Klosik and S. Bornholdt, PLOS ONE **9**, 1 (2014).

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

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

[30] V. Larivière, É. Archambault, and Y. Gingras, Journal of the American Society for Information Science and Technology **59**, 288 (2007).

[31] This assumption is for simplicity, but it is grounded in the reality: The number of citations an article receives spikes few years after its publication [28].

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

[33] Physics Today (2014), 10.1063/pt.5.2012.

[34] R. K. Pan, S. Sinha, K. Kaski, and J. Saramäki, Scientific Reports **2** (2012), 10.1038/srep00551.

[35] A. L. Porter and I. Rafols, Scientometrics **81**, 719 (2009).

[36] E. Garfield, Science **122**, 108 (1955).

[37] A. F. J. van Raan, Scientometrics **62**, 133 (2005).

[38] K. W. Boyack and K. Börner, Journal of the American Society for Information Science and Technology **54**, 447 (2003).

[39] D. W. Aksnes, Journal of the American Society for Information Science and Technology **57**, 169 (2005).

[40] P. O. Seglen, BMJ **314**, 497 (1997).

[41] E. Favaloro, Seminars in Thrombosis and Hemostasis **34**, 007 (2008).

[42] B. S. Frey and K. Rost, Journal of Applied Economics **13**, 1 (2010).

[43] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Rev. Mod. Phys. **87**, 925 (2015).

[44] C. Castellano, S. Fortunato, and V. Loreto, Rev. Mod. Phys. **81**, 591 (2009).