Parolo, Pietro della Briotta; Kujala, Rainer; Kaski, Kimmo; Kivela, Mikko

Tracking the cumulative knowledge spreading in a comprehensive citation network

Published in: PHYSICAL REVIEW RESEARCH

DOI: 10.1103/PhysRevResearch.2.013181

Published: 20/02/2020

Document Version
Publisher's PDF, also known as Version of record

Published under the following license: CC BY

Please cite the original version:
Parolo, P. D. B., Kujala, R., Kaski, K., & Kivela, M. (2020). Tracking the cumulative knowledge spreading in a comprehensive citation network. PHYSICAL REVIEW RESEARCH, 2(1), [013181]. https://doi.org/10.1103/PhysRevResearch.2.013181

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.
I. INTRODUCTION

Since the seminal work of de Solla Price [1] quantitative analysis of knowledge spreading through a network of scientific publications has become a matter of great interest. The analysis of bibliometric data not only sheds light on the structure of science and its knowledge accumulation, but also gives us insight into the citation distributions [2–4], collaboration networks [5,6], geographical patterns of collaborations and citations [7–9], and the structural changes that take place at the level of scientific fields [10–12]. Along this line of research, the citations between scientific publications are in the focus of interest and they can encode various meanings between publications [13], but perhaps most often they indicate that some knowledge from the cited publication is being used in the citing publication [14]. Despite all this progress for more than half a century, a core question remains elusive: at the global scale, where is the knowledge going and where is it coming from?

The main research paradigm in the Science of Science has been to focus locally on the direct citations between a pair of publications. This thinking is exemplified by the literature on quality measures that are based on direct citations, such as the H index [15], the journal impact factor [16], and a number of others. Even though much attention has been given to the structural limitations of these methods [17–19], the standard approach to overcome such limitations has been to introduce minor adjustments while still relying on the numbers of direct citations each publication/author/journal receives [20–23]. This local paradigm is in contrast to the structure of science itself because science is a cumulative process where researchers “stand on the shoulders of giants,” i.e., the results of each researcher are intrinsically based on a massive amount of previous work, not just the publications that are directly cited [14,24]. Therefore, when one attempts to study the structure and behavior of scientific knowledge accumulation, it is necessary to look at the whole process and not focusing only on a local area of the system.

In this work, we aim at answering the following question: starting from a publication or a group of publications, where and how does its knowledge flow in a citation network if one looks beyond the direct citations? To answer this question, we study all the possible chains of citations, indicated by the models of persistent influence and diffusion. These stylized models are complementary to each other and are computationally tractable even if they are applied to the citation network of all the scientific publications. In particular, we introduce two stylized models for studying the flow of knowledge, i.e., the models of persistent influence and diffusion. These stylized models are complementary to each other and are computationally tractable even if they are applied to the citation network of all the scientific publications. Here, we use these models to study a comprehensive data set of around 35 million publications from most fields of science covering more than hundred years of making research (see Appendix A for details).
When we use millions of individual publications and different groups of them as the sources of information chains, we obtain a good general understanding of the knowledge flows in the citation network. When we compare the relationship between the direct citation counts and the total persistent influence of individual publications, we find that papers associated with Nobel prizes tend to outperform their peers with similar citation counts and publication years. Given that papers related to Nobel prizes are expected to have a profound long-term impact on science, this result validates in part our global approach to model the spreading of scientific knowledge.

There are some previous works looking at the impact of citations beyond the local perspective, but often it was done from a very different starting point as compared to the present work. The spread of information has been studied through the adoption rates [25], contagion models [26], and diffusion [27,28], but these studies have either relied on small samples of the full citation network or of the aggregated networks. This is in contrast to our approach to model the processes on a comprehensive network of millions of publications, which allows us to track individual publications and where the processes are independent of any sampling or categorisation of journals and publications. Further, the PAGERANK-type algorithms have been used to rank publications [29] and individual scientists [30] in a set of physics journals. These PAGERANK-type methods are closely related to diffusion but they rely on random walkers that do not keep track of their origins and destinations, whereas we are interested in how the information from one publication (or collection thereof) is used in other publications.

In addition, it is possible to quantify the spreading of scientific ideas, or memes, between citing and cited publications [31]. Perhaps the closest work to ours is the study of in-components of individual publications in a citation network of physics papers [32], which has enabled to pinpoint influential, components of individual publications in a citation network of publications, which allows us to track individual publications and where the processes are independent of any sampling or categorisation of journals and publications. Further, the PAGERANK-type algorithms have been used to rank publications [29] and individual scientists [30] in a set of physics journals. These PAGERANK-type methods are closely related to diffusion but they rely on random walkers that do not keep track of their origins and destinations, whereas we are interested in how the information from one publication (or collection thereof) is used in other publications.

This paper is organized as follows. First we introduce the persistent influence model and use it to track the amount of influence the millions of individual publications have on all other publications downstream to them in the citation data we use. After summarising these results, we explore how the papers associated with a Nobel prize perform in terms of persistent influence, and explore the publications whose rankings in terms of direct citations and persistent influence differ the most. Finally, we introduce the diffusion process and focus on the rate of diffusion out of scientific fields, subfields and journals, and how their speeds have changed over the years.

II. PERSISTENT INFLUENCE PROCESS

Our aim in this study is to model how the scientific knowledge percolates through the network of publications citing each other. Since the flow of knowledge within scientific publications is difficult to measure or quantify using available data, some simplifying assumptions are required. First, we assume that each publication is only using information that is present in the publications it cites, and the amount of intrinsic information it contains is negligible to it. Second, we assume that each publication contains the same amount of information. Third, we assume that each of the cited publications is equally important for the citing publication. Fourth, we assume that the information content of a publication can be presented as the weighted sum of information contents of the different sources, as opposed to some more complicated function.

The above ideas are formalized in a simple persistent influence spreading process. Starting from a seed publication \(s\), we attribute to it an initial value of influence \(I_{s \rightarrow s} = 1\), while all the other publications have an initial value of 0. We then compute the influence of the seed publication on publications published after the seed by going through the publications in chronological order. Now, each publication can be directly influenced by the seed or inherit the influence of the seed through longer chains of citations. Assuming that each of the cited publications is equally important and that the influence can be summed, the rule for calculating the persistent influence of the seed publication \(s\) to publication \(j\) is

\[
I_{s \rightarrow j} = \sum_{i \in I^m_j} \frac{I_{s \rightarrow i}}{k^m_{i,j}},
\]

where \(I^m\) is the set of publications cited by \(j\), and \(k^m_{i,j} = |I^m_j|\) is the number of references, or the in-degree, of publication \(j\). The normalisation guarantees that the sum of influence that the cited publications can have on publication \(j\) is at most 1 in the case that all the cited publications also have persistent influence values of 1. For an illustration of this model, see Fig. 1.

In the presented model, each citation in the reference list of a publication is considered equally important. This feature of the model has some consequences that are important to understand. 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. Similarly, a publication that is cited by a review publication that 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 publication is influencing the review.

When the process continues beneath the first few layers of publications in the chains of citations, typically a large number of papers become influenced by the seed node while the persistence influence values for individual papers become diluted. Now, instead of studying the persistent influence of the seed node to an individual paper, it is more meaningful to study the persistent influence of the seed publication to a group of publications. For instance, given a seed publication, one could compute an aggregated persistent influence value \(I_{\text{Physics}}(t = 2019)\) by summing over the persistent influence values of all the publications published in physics journals during the year 2019.

\[^1\]This assumption could be relaxed by introducing a parameter describing the fraction of new information in each new paper, which would act like a damping factor. For simplicity, here we only study the case where this fraction is zero.
A. Persistent influence in a random citation network

One can gain intuition on how the spreading process of persistent influence works by considering how the papers’ influence scores develop in a simple model of a citation network. Let us consider a citation network where publications are published in generations, and they only cite publications of the previous generation. Further, we assume that the number of publications $n(t)$ in each generation $t$ grows at a constant rate $\mu = n(t+1)/n(t)$, and the maximum values of the in- and out-degree distributions are always small compared to the current system size $n(t)$. Now, if the in- and out-degrees of papers are statistically independent, the sum of the influence values of all the papers in generation $t+1$ is on average given by

$$I(t+1) = \mu p(k^m = 0) I(t),$$

where $p(k^m = 0)$ 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 grows or decays exponentially with the factor $\mu p(k^m = 0)$ or remains constant if the generation sizes do not change and there are no “dead-end” publications, or if these two factors counter each other out. In reality, the scientific input has been continuously growing ($\mu > 1$) [33], and thus we expect that on average the influence of early papers will grow. However, this is only the average picture. While some papers’ influence values die out, others’ grow faster than the average.

B. Empirical results

We will now apply the process of persistent influence to a citation network with millions of publications covering most fields of science. The data also contain information about the journal each publication is published in and the subfield classifications of the journals. These data are combined with a coarser field classification of the subfields and the information about the papers that led to Nobel prizes [33], see Appendix A for details. We begin by examining the persistent influence profile of a publication depicting how we can evaluate the cumulative influence of a paper on other papers, journals, and fields of science across time. Later, we move on to a large-scale analysis of source publications and assess the outperformance of papers related to Nobel prizes.

1. Case study: Glauber’s Nobel winning publication

Roy J. Glauber’s seminal paper on photon correlations [34] was published in 1963, which eventually led him winning the Nobel Prize. Around nine years after the publication date, already 10 thousand publications can be connected to Glauber’s paper through a chains of citations. The subnetwork containing these publications and the citations between them are visualized in Fig. 2(a). Different subfields in this network of persistent influence can be seen to roughly organize to their own branches, and the developments within these subfields can be approximately assessed. For example, publications in interdisciplinary physics journals are seen to be the early ones citing Glauber’s work, but after few years the publications in the sphere of influence are more and more in the journals categorized to more specific fields of physics.

As expected, the number of publications that could be indirectly influenced by Glauber’s paper in the citation network grows exponentially, reaching millions of publications before 2008 that is already a significant part of all the publications in the data set (compared to 473 direct citations). The size and topology of such a large subnetwork may be of little practical interest, but the persistent influence values are still meaningful: even though a massive number of publications could be influenced through chains of citations on long time ranges, most of the persistent influence values are extremely small. That is, even for the long time ranges, the influence is concentrated on a small subset of publications that can often be reached through multiple routes. By aggregating yearly influence values within the categories of publications, we can see in Figs. 2(b), 2(d), 2(f), and 2(h) that these publications are mostly on fields and subfields which are known to be impacted by Glauber’s work.
FIG. 2. Persistent influence profile of Glauber’s Nobel prize winning paper from 1963 to 2008. (a) The first ten thousand, or about 9 years of, papers and citations between them in the downstream network of Glauber’s paper. The edges are colored according to the subfield of the citing paper. In case of multiple subfields, the color is chosen according to which of these subfield has the largest area in (d). Glauber’s paper is at the center of the figure. The polar coordinates of the other publications are chosen so that older publications are closer to Glauber’s paper, and the angles are determined using a force-directed layout algorithm. (b), (d), and (f) show the relative, aggregated persistent influence values of Glauber’s paper on different groups of publications based on the publication year. (c), (e), and (g) show similar yearly sums for local influence values that have been calculated using only direct citations. (b) and (c) show the relative distribution of influence among fields (F) of science ordered in decreasing order of area in the figure starting from the bottom. The most influenced fields are physics (green), mechanical (orange) and engineering (light blue). Similarly, (d) and (e) show the influence on subfields (SF) with the most influenced subfields being optics (red), physics, and multidisciplinary physics, and (f) and (g) show the influence on journals (J) with Physical Review A (light green), Physics Letters A (light purple), and Physical Review Letters (light orange) being the most influenced ones. Only the contribution of eight largest fields/subfields/journals are shown and the rest is shown as light grey space. In (g) and (h), we show the total influence at the resolution of one year, i.e., the sum of influence values $I(t)$ for publications published in each year. The legend on the bottom corresponds to subfields, that is, (a), (d), and (e).

To illustrate the difference between the local viewpoint of counting direct citations and the persistent influence, we show a local influence profile calculated using Eq. (1) but only considering direct citations to the Glauber’s original publication in Figs. 2(c), 2(e), 2(g), and 2(i). At first sight, the most notable difference between the global and local profiles is that the global profile is much smoother than the local one. This is expected as if a field was influenced by a seed paper, the papers from the field are also likely to cite the seed. However, a closer inspection reveals more subtle differences between the global and local profiles. The influence of Glauber’s paper on publications categorized to fields [Fig. 2(b)] shows a strong persistence in physics with a gradually growing contribution to fields called mechanical and engineering. This pattern is already stable after ten years in the persistent influence, but it is not visible in the local profile even though the second most influenced field, mechanical, starts to cite the seed publication later on. A similar effect of optics becoming prominent in the persistent influence much before it is picked up by the local influence profile can be seen when looking at the categorisation in the level of subfields [Figs. 2(d) and 2(e)]. Going down to the level of journals we can see that the contribution to optics is mainly due to two journals: Physical Review A and Optics Communications, which once again are not as strongly present in the local profile [Figs. 2(f) and 2(g)].

The total persistent influence of Glauber’s publication on publications published during each year grows (except for few years), but the local influence remains relatively stationary when one considers only direct citations. This observation is qualitatively in line with what would be expected based on the generational model defined in Sec. II A, because the persistent influence values are sensitive to the expansion in science that has happened at a rapid pace since 1963. However, it is not clear from this analysis if the rate at which the persistent influence grows is a typical example of a publication of that time or if this Nobel winning publication is somehow special.

2. System level analysis of seed publications

In order to gain insight into the relationship between the persistent influence and local citation count at the system level, we repeated the influence profile calculation described
total persistent influence values of publications that receive similar numbers of citations. This dependence between the local citation counts and the globally computed persistent influence values resembles the results in Ref. [29] obtained for another global measure of publication importance. Further, this result indicates that the number of citations per se is not sufficient to fully summarize the persistent influence that a single paper has had within the scientific literature. Also, we see that older papers manage to gather a significantly higher amount of impact with the same number of citations. This is expected, as the older papers have had more time to gather cumulative influence among their scientific off-springs, and may also benefit from the growing system size.

To remove this advantage of the older papers, we calculate the total persistent influence that each paper gathers during the first 10 years since its publication date \( t_0 \) \( \langle I_{t_0} \rangle \). We have chosen to use 10 years as the limit so that there would be sufficient time for publications to gather citations and persistent influence. Note however, that if 5 years were used as the limit instead, the results would remain qualitatively similar. This partly removes the effect of the publication year, but older publications with the same number of citations are still seen to have slightly higher total persistent influence as compared to the newer ones [Fig. 3(b)].

The remaining difference between the publication years might be due to other changes in the citation practices such as the increase in the length of reference lists, which causes the denominator in Eq. (1) to reduce the amount of influence to a single citing paper. However, the distributions of persistent influence values of publications with the same number of citations within 10 years \( \langle C_{10} \rangle \) have superexponential tails that are relatively independent of the publication year [Fig. 3(c)]. This shows that even on a time scale of 10 years, publications having similar numbers of citations can have varying scientific impacts when measured using the total persistent influence.

3. Performance of Nobel-winning publications

What are the publications that achieve higher persistent impact values than expected based on their citation count and are these publications somehow more important for science than the others? To begin to answer these questions, we studied 74 papers that were associated with a Nobel prize [35], which we assume here is a hallmark of scientific importance.

The Nobel prize winning papers mainly populate the top-right corner in Figs. 3(a) and 3(b), indicating that, generally, they tend to gather high absolute values of citations and influence. However, as many of them are among the few most cited papers in the data their persistent influence is not significantly higher than other papers with similar citation counts. Because the Nobel prizes are often given with a significant and growing delay [35] indicating that their significance is acknowledged only long after the original research is published, we divert our attention to the longitudinal aspect of the influence process. The idea behind this approach is to take a set of control papers which, at a certain point in time, are apparently equivalent to the Nobel paper and to compare the performance of the Nobel paper with respect to the control group at a later point in time. Practically speaking, we construct a set of control papers that have been published in the same year, \( t_0 \), as the reference...
paper and have at most 10% difference in the citation count after 10 years, $C_{10}$. We calculate the out-performance of a Nobel paper to its controls as the ratio $I_I/C_{I,C_{10}}$, where $I$ is the total persistent influence value of the Nobel winning publication and $C_{I,C_{10}}$ is the average total persistent influence of the control papers. These outperformance values can be compared to the out-performance of Nobel papers in terms of total number of citations $C_{I,C_{10}}$. Out of 74 Nobel papers 65 outperform their controls in persistent influence and 50 of them outperform the controls in citations [see Fig. 3(d)], which makes the persistent influence significantly more likely to display outperformance than citation counts ($p$ value $< 10^{-4}$ for the hypothesis that $50/74$ chances could have produced 65 or more successes). Furthermore, in 73% of the cases (54 papers out of 74) the persistent influence outperformance is greater than the citation one (again, the $p$ value $< 10^{-4}$ for the hypothesis that this difference was produced with a null model that gives equal probability for the persistent influence and citation count to outperform the other).

Based on the above numbers and the overall picture presented in Fig. 3(d), it can be said that the Nobel papers outperform the controls in terms of persistent influence and citation counts, and that typically the outperformance of the Nobel papers is greater in terms of the persistent influence than citation counts. For some Nobel papers, the outperformance is smaller in terms of persistence influence but the difference to citation-count-based outperformance is then typically small. An interpretation of these results is that the Nobel papers often constitute the starting points of new growing areas of science that reach beyond the publications directly citing them, as one would expect from publications with influential results. This conclusion supports the finding of Ref. [32], where the groundbreaking papers by important Nobel laureates were found to have large networks of child nodes spanning over multiple layers.

### 4. Using rankings to detect influential but low-cited papers

The raw values of the persistent influence and citation counts have fat tails, which can make the analysis very sensitive to outliers, and also depend heavily on the publication year. Instead of looking at these raw values, we will now use ranks as a more robust measure of publication importance. To this end, we order the papers published within the same year according to the citation count and persistent influence, and give each paper ranks $R_C$ and $R_I$ in terms of the citations and persistent influence, respectively, such that small rank means high value. We analyze the ranks by the publication year to avoid the biases due to higher values of citation counts and persistent influence for older papers. Now, we are primarily interested in identifying those publications that have high persistent influence values but have received relatively few citations over the years. To this end, we define the relative change in ranking as

$$\delta = \frac{R_C - R_I}{R_I}. \quad (3)$$

Here we divide the absolute difference in rankings by $R_I$ to emphasize the publications that have low ranks in terms of the persistent influence: given two publications with the same absolute difference in ranks, the higher $\delta$ is given to the publication with more persistent influence.

Figure 4(a) shows the cumulative distribution of $\delta$ values at different times for the 74 Nobel papers, a control group of papers with 3% of citations of each Nobel paper and uniformly randomly selected papers. The Nobel publications are seen to be much more likely to be among the publications that have high persistent influence but relatively low citation count than publications in either of the control groups. This effect persists when we use the 10-year persistent influence values $I_{10}$ and citation counts $C_{10}$ for computing the ranks [Fig. 4(b)].

The Nobel publications are overrepresented in the publications with high $\delta$ values, but what are the publications with the largest relative difference in the persistent influence and citation counts? Table I shows the publications with largest $\delta$ values for all the papers in our data set. Some of these publications might be “forgotten beauties” in the sense that they have contributed to the development of several other important publications but were forgotten as these more prominent publications gathered all the citations related to the breakthroughs that were made. Indeed, such forgotten publications appear in the list. The publication with the highest $\delta$ score (No. 1) is related to DNA hybridisation techniques and is cited by several early publications in the emerging field of molecular biology, most notably by E. M. Southern’s work on the Southern Blot, a widely-used method for identifying specific DNA sequences from DNA samples. The second, third and fifth are in the reference list of Sanger’s Nobel paper “DNA sequencing with the chain terminating inhibitors,” which had a massive impact on the biological and medical sciences. Papers 5, 9, 15, and 21 are all linked to the identification, classification,
as it was among the few cited works in the famous (at the time) discoveries in the mathematical field of Fractals, blue light emitting diodes. 25 is a small summary of the recent works that led to his Nobel prize for the invention of efficient and, in general, of Material science. 10 is among Amano’s 11 and 30 are linked to the development of carbon nanotubes to the discovery of high-temperature superconductivity, while many papers from physics, with 4, 12, and 24 being linked in this list diverse aspects of science are included. There are virtually the only field present (see Table III in the Appendix), for the highest cited papers for each year where Biology is due to it being the largest field in our dataset. However, unlike AIDS, leukemia).

Finally, the list also shows evidence of the relatively new or prediction of very well known diseases (prostate cancer, AIDS, leukemia).

As expected, Biology is the main contributor to the list, due to it being the largest field in our dataset. However, unlike for the highest cited papers for each year where Biology is virtually the only field present (see Table III in the Appendix), in this list diverse aspects of science are included. There are 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 led 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, as it 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.

Finally, the list also shows evidence of the relatively new field of complex networks, with 23 being among the earliest papers in the field and cited by virtually all the most significant early publications in the field. Overall, we can see how δ is able to describe 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.

### III. DIFFUSION

The persistent influence spreading method we just introduced is a simple and elegant method to model 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 publications 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 publications but across journals, subfields, and fields.

The idea behind the diffusion method is to start a random walker from a seed publication that is randomly selected from a set of seed publications, and then, at each step, let it move from a publication to future publications citing it.
(see Fig. 1). That is, if we have $N$ seed papers we assign the same initial probability $\hat{D}_{s \rightarrow i} = 1/N$ to all of them, and calculate the probability that the random walker goes through the rest of the publications as

$$\hat{D}_{s \rightarrow j} = \sum_{i \in \Gamma^\text{in}_j} \hat{D}_{s \rightarrow i} k_i^{\text{out}},$$

where $\Gamma^\text{in}_j$ is the set of publications cited by $j$, and $k_i^{\text{out}}$ is the number of publications citing $i$ (or, the out-degree).

Note that this process is sensitive to the time window we choose, as the future publications that we do not know about yet will change the degrees $k_i^{\text{out}}$, and thus probabilities of trajectories of the walkers. 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 of each observation year, and the random walk process is recalculated for each starting year and observation year pair. In addition, we are not interested in the raw probabilities of random walkers visiting the nodes $\hat{D}_{s \rightarrow j}$, but on the conditional visiting probabilities $D_{s \rightarrow j}(t)$ given that the walker is within given year $t$. Further, we will disregard any citations within each observation year to avoid emphasising the publications that are published towards the end of each year. It is also possible to define the model such that both the distribution of the initial mass and the diffusion process are based on the citations of the seed papers and of the child nodes. The results, although qualitatively different, are quantitatively identical. See Appendix B for details about these methods and Appendix C for details of how the diffusion probability is calculated.

**Relationship to persistent influence**

The equation for diffusion, Eq. (4), resembles the equation for the persistent influence, Eq. (1), but the similarity between these two runs deeper than that. In fact, the persistent influence values $I_{s \rightarrow i}$ can also be considered in terms of a diffusion process that goes backwards in time. That is, the persistent influence value starting from a single seed publication $s$, $I_{s \rightarrow i}$, is equal to $\hat{D}_{s \rightarrow i}$ in a network where the directions of the links are reversed and the diffusion process starts from the node $i$. In this sense, the two processes, the persistent influence and the diffusion are complementary to each other.

**A. Empirical results**

When starting the diffusion process, we select a field, a subfield, or a journal and a starting time $t_0$, and track the probability that the random walker is found in a field $D_F(t)$, a subfield $D_{SF}(t)$, or a journal $D_J(t)$ at time $t$ (using a resolution of one year). This process is illustrated in Figs. 5(a), 5(c), and 5(e) for the initial field of economics in 1970, [c] and (d)] evolutionary biology in 1980, and [(e) and (f)] the British Medical Journal in 1990. [(a), (c), and (e)] The area in the bottom shows the amount of scientific value retained by the initialization field (F), subfield (SF), or journal (J). The contribution of the eight largest fields/subfields/journals are also shown individually, and the contribution of other fields/subfields/journals are shown as light grey space. In (c), the colors represent the following subfields (from bottom to top): evolutionary biology, biology, miscellaneous, plant sciences, anthropology, ecology, zoology, genetics and heredity, arts and humanities, and general biology. In (e), the colors represent the following journals (from bottom to top): Br. Med. J., Lancet, Br. J. Gen. Pract., Bmj-British Medical Journal, Med. J. Aust., Postgrad. Med. J., Arch. Dis. Child., J. Clin. Pathol., and Med. Clin. (b), (d), and (f) instead show the renormalized value of $D$ retained within each field/subfield/journal for different years (markers) and with the exponential fitting (solid line).

As the diffusion process progresses, the random walker jumps between publications that can be in different groups (i.e., fields, subfields, or journals) with the possibility of returning to the original one. However, as expected in any diffusion process, the random walker will forget its origin, and the amount of probability mass in the original group goes down monotonically as time progresses. This process is observed to be slow in the examples of Figs. 5(a)–5(c). For economics, the probability of finding the walker in other fields has not grown significantly even after 40 years of random walk. Similarly, for the subfield of evolutionary biology and British Medical Journal, the walker is far from forgetting its origin during the observation period. These two initial groups seem to lose probability mass very fast to other groups in the beginning, but this fast phase is then followed by a very slow change, which is almost like a plateau compared to the initial rate of change.
In a stationary system, which the citation network is not, one would expect the diffusion eventually to completely forget its origin and settle to a stationary distribution. Furthermore, such stationary citation system would likely be ergodic, and the stationary state would be unique. The behavior observed for the fields, subfields and journals do not reach such a unique stationary state, but they seem to be rapidly reaching a metastable state, where the probability of finding a random walker in a specific field, subfield or journal does not change much in time but where these probabilities are not independent of the origin. This observation could, for example, be explained by a network structure containing strong clusters where the random walkers get trapped for the time scales of the data.

B. Summarizing the diffusion curves

The above procedure of observing the diffusion patterns is very cumbersome if the goal is to get an idea of the system-wide behavior of diffusion starting from different groups. In order to summarize the results of the patterns such as the ones presented in Figs. 5(a), 5(c), and 5(e), we have looked only at the amount of value retained by each group. For each starting year $t_0$, we observe the yearly values of the initial group $D_C(t)$ and normalize them with the initial value before any diffusion $D_C(t_0)$. We then fit each curve with an exponential of the form:

$$D_C(t)/D_C(t_0) = (1 - \beta)e^{-\alpha t} + \beta,$$

which follows well the typical shape we observe for the curve. The fits for the previously discussed example cases are shown in Figs. 5(b), 5(d), and 5(f). These fits allow us to summarize both the rate of change of the value of the initial years (through $\alpha$) and find the final plateau value (through $\beta$). Therefore $\alpha$ can be used to measure the rate at which one group shares its knowledge with other groups that are close to it, and $\beta$ instead represents the intrinsic “conservativeness” of a group and the other groups related to it, i.e. the amount of knowledge retained within the boundaries of the group itself in medium time scales.

In order to provide an easier metric for the exponential decay, we introduce a related parameter called half-life $t^{1/2}$ defined as the time required to lose half of the possible plateau value. That is by solving

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

we get the conventional definition of half-life:

$$t^{1/2} = \ln(2)/\alpha.$$  \hspace{1cm} (7)

C. $t^{1/2}$ and $\beta$ in the data

With the above-mentioned ideas on summarizing the diffusion processes in mind, we can put together information about all the possible initial times and fields, subfields, and journals. Table II shows the values for the half-lives, $t^{1/2}$, and plateau values, $\beta$s, in 1970 and 1995. We can see that in general there is a decreasing trend for half-lives while the plateau value $\beta$ instead shows a much more stable pattern.

Some of the individual fields display interesting patterns. 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, its change in $\beta$ is positive and the second highest (behind music).

This could indicate that *multidisciplinary* might have become a field of its own, which can retain random walkers within itself for long periods of time. This observation is coherent with the evidence that shows the increasing role of interdisciplinarity in science \[10,11,36,37\]. It is also interesting to note that while in 1970 some humanistic fields show very high values for their half-lives, e.g., *philosophy, history, anthropology, literature, and linguistics*, these fields also show some of the highest changes in time, putting them much closer to hard sciences these days than they were before.

A more systematic observation on the changes in the speed of the diffusion process is shown in Figs. 6(a), 6(c), and 6(e) which displays the change of half-lives for a set of fields and changes in the distributions of half-lives for all subfields and journals. All of the fields show a speeding-up pattern, losing between 20% and 60% of their half-life values, while for subfields and journals the more recent cumulative distributions of half-lives are above the older ones, showing that the values have in general decreased.

D. Renormalizing the time

Previous studies show \[33\] that the time as such may not be the best choice for measuring the rate at which changes happen in science, but instead use the numbers of papers published as the measure of progress. In other words, science can be considered to be “updated” every time a new publication is introduced in the system, and the count of such updates is a better measure of “progress” in science than simply the time elapsed.

<table>
<thead>
<tr>
<th>Field</th>
<th>1970</th>
<th>1995</th>
<th>$\Delta$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Philosophy</td>
<td>19.7</td>
<td>0.84</td>
<td>4.36</td>
</tr>
<tr>
<td>Economics</td>
<td>11.0</td>
<td>0.83</td>
<td>4.20</td>
</tr>
<tr>
<td>Psychology</td>
<td>8.93</td>
<td>0.72</td>
<td>3.44</td>
</tr>
<tr>
<td>Linguistics</td>
<td>8.86</td>
<td>0.87</td>
<td>3.02</td>
</tr>
<tr>
<td>Chemistry</td>
<td>8.55</td>
<td>0.80</td>
<td>1.99</td>
</tr>
<tr>
<td>Music &amp; Dance</td>
<td>7.83</td>
<td>0.82</td>
<td>6.18</td>
</tr>
<tr>
<td>Gen. Humanities</td>
<td>7.25</td>
<td>0.85</td>
<td>3.43</td>
</tr>
<tr>
<td>Mathematics</td>
<td>7.14</td>
<td>0.87</td>
<td>3.21</td>
</tr>
<tr>
<td>Medicine</td>
<td>6.54</td>
<td>0.83</td>
<td>3.20</td>
</tr>
<tr>
<td>Sociology</td>
<td>6.34</td>
<td>0.80</td>
<td>3.72</td>
</tr>
<tr>
<td>Engineering</td>
<td>4.89</td>
<td>0.82</td>
<td>2.33</td>
</tr>
<tr>
<td>Law</td>
<td>4.38</td>
<td>0.92</td>
<td>7.21</td>
</tr>
<tr>
<td>Social Sciences</td>
<td>4.38</td>
<td>0.73</td>
<td>2.35</td>
</tr>
<tr>
<td>Physics</td>
<td>4.01</td>
<td>0.82</td>
<td>2.32</td>
</tr>
<tr>
<td>Management</td>
<td>3.72</td>
<td>0.78</td>
<td>3.60</td>
</tr>
<tr>
<td>Biology</td>
<td>3.43</td>
<td>0.71</td>
<td>1.69</td>
</tr>
<tr>
<td>Multidisciplinary</td>
<td>1.33</td>
<td>0.59</td>
<td>1.08</td>
</tr>
</tbody>
</table>
As the number of publications per year $N(t)$ grows exponentially, one would theoretically expect that the functional form of the exponential decay given by (5) would have to be adjusted when measuring time in terms of published papers. However, as the growth rate of yearly publication number is sufficiently small $N(t) \approx N_0 e^{\delta t}$ with $\delta \approx 0.05 \text{ (year}^{-1})$ across all the fields, we can approximate this exponential growth by a simple, linear function $e^{\delta t} \approx (1 + \delta t)$ for the span of time $t$ that we consider. This allows us to use the functional forms given by Eq. (5) also for fitting the half-lives $t^2$ and the asymptotic fractions $\beta$ when the time is measured in terms of published papers.

Hence we are able to follow the earlier procedure we applied to real time and quantify the half-lives in terms of the numbers of published papers. We chose to use the number of published papers in each field as a unit of time for the fields, while for the subfields and journals we used the data from all scientific publications.

In Figs. 6(b), 6(d), and 6(f), we show the half-lives that are fitted in terms of published papers. Interestingly, now the half-lives do not in general decrease as a function of the publication year as only chemistry shows such a clearly decreasing pattern. The other fields, instead, either remain constant or show a significant increase in their half-lives over time. A very similar result can be seen in the half-life distributions for subfields and journals, which are shifted to higher numbers of papers as the initial time increases. That is, these results suggest that the growth in the speed of diffusion can be explained by the increase in the rate at which publications are published.

IV. DISCUSSION AND CONCLUSIONS

Ever since bibliometric data of scientific publications have been available, there has been efforts to analyze such data with the goal of quantifying scientific research. The dominant framework has been to use the counts of direct citations between publications, journal, and research fields in order to quantify the relationships between them, and, for example, to rank authors [15], publications [38], universities [39], and institutions [40]. As these methods work reasonably well [41], and apart from few exceptions [29,30], the improvements of these methods have been correcting technical flaws [42,43]. However, most of the previous research has ignored the intrinsic conceptual issue of methods based on just counting citations: they ignore the fact that scientific publications are not only based on the information created in the publications they cite, but on the whole body of literature underneath this first layer of publications.

We have introduced two simple methods to analyze how the knowledge created in a publication, or in a group of publications, might percolate through the scientific literature. This approach follows the tradition of modeling the dynamics on networks, where a real observed network is used as a substrate where the progress of a stylized model is tracked. This approach has been extensively used in network science to study epidemic spreading [44] and social dynamics [45]. In all of these cases, the models are not expected to exactly mimic the real behavior, but the goal is to reproduce the behavior in the large scale with an accuracy that is enough to make at least qualitative statements about the system. Our goal in this work was to explore this approach in knowledge spreading in citation networks.

The first of our two measures, persistent influence, is based on the idea of papers inheriting the knowledge of papers they cite, i.e., the “shoulders” on which they stand. As a consequence, a paper can influence later papers through chains of citations. As expected, the out-degree, in other words the direct citation count, is positively correlated with the persistent influence, but we also observe that papers that are similar in publication date and citation count can have a wide range of the persistent influence values. In our simulations, we have found publications that have several orders of magnitude higher influence values than papers with similar numbers of citations. This finding suggests that there are publications for which the citation count can be a poor proxy for tracking the global cumulative influence.

We tested the hypothesis that the discrepancies in the citation counts and global persistent influence values are not meaningful but simply noise added by the global process, and to do this we used papers associated with Nobel prizes as a manually curated corpus for presumably high influence on science. We found a significant overperformance in terms of the persistent influence by the Nobel papers when compared to the publications with similar initial citation counts and publication dates. Thus the indirect influence (i.e., the
persistent influence) and direct influence (i.e., the citation count) seem to capture different aspects of the influence of a publication. 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 the fields that would later become hot topics of their time in the scientific world.

The second modeling approach we employed was a simple diffusion method. We focused on analyzing the rate of diffusion of knowledge across the fields, subfields, and journals. We summarized the curves describing the loss of diffusing knowledge to other fields, subfields, and journals by an exponential decay function that reaches a plateau value. Each starting time and set of seed publications can thus be described by a plateau value of the retained knowledge $\beta$ and by a typical time required $t^{1/2}$ to share half of the available knowledge. We found that $\beta$ 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 the interdisciplinarity of research. However, we showed that the faster sharing of information could be explained by the increase in the rate at which publications are produced.

One crucial difference between the diffusion and persistence influence models is that in the diffusion model the total amount of knowledge created in some publication is conserved across papers citing it, whereas in the persistent influence model the knowledge is allowed to “duplicate” across papers. Given that there is no reason why many publications could not contain a same piece of knowledge, out of these two models the persistent influence approach is likely to provide a more realistic picture of how knowledge spreads in citation networks. The diffusion on the other hand could be considered as more of a theoretical tool to understand the structure of the citation networks.

The work done here forms a basis for future possibilities of the model-based approaches to track global knowledge spreading in citation networks. For example, more detailed look at the long-term destinations of influence starting from various sources could bring interesting results. Note also that nothing would stop one, to introduce the initial persistent influence to a group of papers (instead of a single publication) similar to the diffusion process, and repeat the type of analysis done here for fields, subfields and journals using the persistent influence model. Furthermore, one can easily reverse the tracking direction of the persistent influence model and investigate which publications, or groups of publications, in the history have influenced individual papers. One can also make the influence spreading more realistic with the cost of increasing the complexity of the model. For example, the amount of knowledge created by each publication can be added as a parameter, which will effectively work as a damping factor that will decrease the influence of very long chains of citations.

We have introduced relatively simple methods to analyze the spreading of knowledge in citation networks at a global scale, and shown that these methods can lead to significantly different results than what can be obtained by using the local approach. With more and more bibliometric data being available, we hope that our findings will 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.

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 EU HORIZON 2020 INFRAIA-1-2014-2015 Research Infrastructures SoBigData: Social Mining, Grant Agreement No. 654024. 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: DATA DESCRIPTION

We use the data set that consists of all publications (publications and reviews) written in English from the year 1898 till the end of 2013, included in the database of the Thomson Reuters (TR) and now Clarivate Analytics’ 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 publications and journals for which this 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 as in Ref. [33]. The set of papers associated to the Nobel prize is identical to the one described in Ref. [35], which is a manually curated data set constructed using the Nobel Foundation website as a primary source. Note that Nobel prizes are not explicitly given to individual papers, and there can be some ambiguity in determining the papers associated to the prizes [46].

We use the set of citations between the filtered publications to construct a network where there is a link from citing publication to the cited publication. We use the publication time information to remove links where the date of the cited publication is equal or later than the date of the citing publication. 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 publications, for which most publications citing them are not in our data set, we only consider the nodes in the citation network until the year 2008. Previous literature [47] 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 publications published until 2008 minimizes the boundary effects originating from the missing data on future publications.

APPENDIX B: ALTERNATIVE DIFFUSION METHODS

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): $D_i = \frac{\sum_{t=1}^{\min(\text{year}-2008,5)}}{\sum_{t=1}^{\text{year}}}$. 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 publication to each citing publication with the same probability, and in the other case the random walker preferentially chooses publications that receive more citations in the coming five years $C_i(t < 5)$. This is equivalent to setting the weight $w_{ij}$ of each link from paper $i$ to a citing paper $j$ to one of the two values:

$$w_{ij} = \frac{1}{\sum_{l \in \Lambda_1} 1} = \frac{1}{k_{i\text{out}}} \quad \text{or} \quad w_{ij} = \frac{c_i(t < 5)}{\sum_{l \in \Lambda_1} c_l(t < 5)}.$$

The results for both processes are similar and here we only show them for the simpler process. For technical details of how the process was made computationally tractable and implemented see Appendix C.

APPENDIX C: 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 Description section. After this it was necessary to order the nodes in topological order, which guarantees that for every directed edge connecting paper $i$ (the citing paper) and $j$ (the cited paper), $j$ comes before $i$. Such ordering should, in principle, correspond with the time stamp of the publication. However, for older papers, the topological ordering of the publications could not be resolved using publication dates, as only year information was available. Therefore we took advantage of the fact that papers published in earlier years are bound to have a lower rank in the order. Thus we built yearly citation networks and ordered them topologically, starting from the oldest one. 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. When looping through the topologically sorted edge-list, this ordering guarantees that each paper has collected all the value/influence upstream before pushing its own forward, and that all value/influence contained by the paper is pushed forward to the citing papers at once. Further, this ordering allowed us to loop through the edge-list 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 $y_{\text{start}}$. 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 $y_{\text{end}}$ as the year in at which we will stop pushing values forward. Hence we loop through the node-list of all scientific papers in our dataset published between $y_{\text{start}}$ and $y_{\text{end}} - 1$ arranged in topological order to initialize the weights for each node. We consider only links to neighbors that point to papers published before $y_{\text{end}}$.

Once the weights for the diffusion process have been initialized we can push the value of each node by looping again through the node-list in by topological order, which guarantees that no value is ever pushed from a node before the same node has collected all the previous values available. The pushing starts from $y_{\text{start}}$ and stops in $y_{\text{end}} - 1$ but spreads to papers published all the way to $y_{\text{end}}$, without pushing any value within $y_{\text{end}}$. 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 its value from another paper. In case of the existence of a sink (i.e., an uncited node) in previous years, we also consider it being part of the final year. However, only a small fraction of nodes remain uncited until 2008, with most papers being only temporary sinks. This means that the almost all of the value is pushed along, with little risk of it being trapped in sinks along the way.

After the pushing has ended we can collect all the values that are left not 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 storing permanently the value of any node that appears in the first column of the edge-list as by definition they will necessarily get rid of all their values. Also, by construction the sum of the values of all the leaves equal 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 $y_{\text{end}}$ between those years, since the network initialized is different each time. This means that when we collect the data in a certain year, we do not 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.

FIG. 7. An example of the diffusion process and the collection of data. (a) shows the original DAG, while (b)–(d) show the result of the diffusion process for each year (delimited by the dotted line). The sum of the values in each red box sums up to one. Note how in the last panel a sink exists and is included as being part of the set of papers for which to sum the values.
factor. The data collection, like the data initialization, can be done on paper, journal, subfield, and field levels.

In Fig. 7, we show an example of the diffusion process and the data collection. We can see that the method explained the previous paragraph causes the intermediate values (e.g., the values of the first time interval in Fig. 7) to be modified when the process is run for more recent years. This is because, new publications and their links to older links cause a change in the link weights computed using Eq. (7).

APPENDIX D: HIGHEST CITED PAPERS

The Table III shows the papers with highest numbers of citations for each year between 1970 and 1999. Biology is virtually the only field present in this list.

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