Knowl Inf Syst (2018) 54:487–509
https://doi.org/10.1007/s10115-017-1080-y

REGULAR PAPER

Universal trajectories of scientific success

Tanmoy Chakraborty[1] · **Subrata Nandi**[2]

Received: 1 September 2016 / Revised: 29 May 2017 / Accepted: 30 June 2017 /
Published online: 12 July 2017

Abstract Success of a scientific entity generally undergoes myriad vicissitudes, resulting in different patterns of *success trajectories*. Understanding and characterizing the rise and fall of scientific success is important not only from the perspective of designing new mathematical models but also to enhance the quality of various real-world systems such as scientific article search and recommendation systems. In this paper, we present a large-scale study of the subject by analyzing the success of two major scientific entities—papers and authors—in Computer Science and Physics. We quantify "success" in terms of citations and in the process discover six distinct success trajectories which are prevalent across multidisciplinary datasets. Our results reveal that these trajectories are not fully random, but are rather generated through a complex process. We further shed light on the behavior of these trajectories and unfold many interesting facets by asking fundamental questions—which trajectory is more successful, how significant and stable are these categories, what factors trigger the rise and fall of trajectories? A few of our findings sharply contradict the well-accepted beliefs on bibliographic research such as "Preferential Attachment", "first-mover advantage". We believe that this study will argue in favor of revising the existing metrics used for quantifying scientific success.

Keywords Scientific success · Success trajectories · Scientific entities · Citation

A part of the research was done when the author was at University of Maryland, College Park, USA.

Electronic supplementary material The online version of this article (doi:10.1007/s10115-017-1080-y) contains supplementary material, which is available to authorized users.

✉ Tanmoy Chakraborty
tanmoy@iiitd.ac.in

Subrata Nandi
subrata.nandi@gmail.com

[1] Department of Computer Science and Engineering, Indraprastha Institute of Information Technology, Delhi (IIIT-D), New Delhi, India

[2] Department of Computer Science and Engineering, National Institute of Technology, Durgapur, India

1 Introduction

What does "success" of a scientific endeavor look like? What are the factors contributing to the rise and fall of scientific success? How do we quantitatively capture the impulsive nature of success? Answering all these questions will lead to unfolding several fundamental characteristics pertaining to the "science of science" [15], which might in turn be used in various decision making processes such as hiring [32], promotion, decisions about funding [6,7] or ranking of individuals and universities [37].

It should therefore be clear that measuring scientific success is important—but it is also elusive. Citations are often regarded as "units of credit" used to pay homage to pioneers (when citing insightful works), or to peers (when recognizing related work), in addition to identifying important methodologies and equipment [20]. There has been a plethora of research focusing on understanding the fundamental mechanism driving the citation dynamics, dating all the way back to the pioneering work of Price in 1960 [15]. Researchers use citations to identify seminal papers, unfold the evolution of disciplines [50], understand team formations [27,52], and propose several bibliographic measures to quantify scientific impact [23,30]. An important issue that had traditionally been ignored in the relevant literature is that the acquisition of citations is time-dependent. To address this issue, several age-based citation growth models were proposed along with measures such as "fitness", "quality" and "perceived novelty" [28,53]. However, such models are incapable of explaining the diverse nature of *citation growth*. They are, for instance, unable to explain papers with delayed recognition (often referred to as "sleeping beauties" in science) [25,51], or papers with frequent and irregular recognition.

There are a few studies addressing the phenomenon of "late awakening" of papers. Garfield [21,22] was the first to provide examples of articles with delayed recognition. Later, Glänzel et al. [26] estimated such delayed recognition and unfolded interesting characteristics of this phenomenon. van Raan [51] first coined the term "sleeping beauty" to refer to delayed recognition of papers. Redner [45] analyzed large Physics datasets and discovered revived classes of articles which were recognized some time after their publication dates. Ke et al. [31] introduced a parameter-free measure to quantify delayed recognitions. However, all the aforementioned studies were entirely focused on delayed recognition, eschewing other types of recognition patterns.

On the other hand, techniques for measuring the quality and perceived success of research based on bibliographic evidence are debatable at best and only recently evolving across many fronts [4,48,54]. Although there has been a consistent effort devoted to understanding the impact of journals and scientific articles [14,16,24,41,42], little has been explored of the scale of individual authors [40]. Most of the author-centric measures, such as h-index [30] and g-index [18], capture either growth or saturation of the scientific success of authors. They therefore fail to capture the *decline of success*. Analyzing the decline of success is important for unfolding several author-centric aspects of research, such as whether an author is still active in the community, how "worthy" her recent publications are, or whether her older papers stand the test of time.

In our earlier work [10], we introduced the idea of various citation profiles of scientific articles in Computer Science domain. We further showed that none of the existing growth models such as Preferential Attachment models [1] can not capture these profiles, and hence we proposed a new citation growth model to mimic these diverse citation profiles. In the following work, we showed how one can use this profile information to predict the future citation count of an article at the time of its publication [9]. We proposed a two-stage stratified

learning framework which in the first stage uses a rule-based approach to map the citation profile of the examined paper to one of the categories; then in second stage the model is trained on papers belonging to only the mapped category to predict the future citation count of the examined paper. We also quantified the interdisciplinarity of a paper (vis-a-vis a domain) by analyzing the citation distribution and contextual properties of papers such as keywords, topics [8].

However, in our previous studies there was no author-centric analysis, and the analysis of citation profiles was only conducted on a small Computer Science dataset [13]. Moreover, a detailed understanding of the possible causes leading to such diverse profile patterns was also missing.

In this work, we focus on understanding the success of two scientific entities—papers and authors—simultaneously, by analyzing their history over a number of years. To that end, we gather and study two massive datasets of publications related to Computer Science and Physics. The *Scientific success* of an entity in any particular year is measured by simply considering the average number of citations received by the entity in that year. Following this, heuristics are employed to understand the temporal growth pattern of success (henceforth referred to as "success trajectories"). Interestingly, we discover six entirely disjoint success trajectories of scientific entities prevalent in both datasets—*Early Risers*, *Late Risers*, *Frequent Risers*, *Steady Risers*, *Steady Droppers* and *Others*. A thorough analysis of these trajectories reveals several interesting phenomena, all of which have been unexplored due to the lack of a large-scale systematic analysis. A few of these observations are in sharp contradiction with some well-accepted beliefs about the "success" of published research. For instance, a general consensus regarding the citation growth of a paper is that a paper receives a lot of citations within the first 5 years of its publication, followed by a decline over the rest of its lifetime (which is equivalent to our "early-rising" pattern). Based on this observation, popular bibliographic metrics such as the *5-years journal impact factor* have been proposed [23]. However, our discovered trajectory patterns may raise questions regarding the appropriateness of these metrics. Another belief that is challenged by our findings is the idea of "first-mover advantage", that states if a paper does not receive citations in the early stage of its lifetime, it will never get cited later [39]. Interestingly, we observe that the entities following late-rising patterns are essentially the most successful in terms of cumulative citations.

We further demonstrate that empirical observation of success trajectories cannot be reconciled with traditional growth models such as the Preferential Attachment model [1] that are based solely on cumulative citations. In order to answer which trajectory is most successful, we observe that Steady Risers exhibit highest impact; "early rise" does not necessarily predict long-term success; self-promotion through frequent self-citations may lead to complete decline of one's success. Furthermore, the stability analysis of these trajectories leads to the conclusion that predicting long-term success is hard with only a few early years of information. The reason is that besides steady droppers, no other category exhibits such strong signals in its early days. This result once again questions the accuracy of recent work by Wang et al. [53] aiming to predict long-term scientific impact and calls for more sophisticated models that are able to capture the diverse range of scientific trajectories. A detailed analysis explaining the reasons behind the rise and fall of trajectories reveals that late-rising papers mostly correspond to "premature" research which are triggered by papers of different disciplines after a long time. However, once acknowledged, these papers have tremendous potential to bring about a completely new line of research. Our study agrees with earlier findings, according to which collaborative and multidisciplinary research tends to break the most ground [1].

To the best of our knowledge, this detailed large-scale analysis of scientific success for both authors and papers together is the first in the field and has the potential to open up numerous research directions. We believe that the key observations made in this paper will help others rethink/reformulate existing bibliographical findings, and to define the dynamics controlling the "awakening process" of scientific entities, shedding light on the reasons why some papers follow unconventional population trajectories.

2 Materials and methods

2.1 Massive bibliographic datasets

We analyze two massive bibliographic datasets to validate the universality of success trajectories: (i) `CS`: we crawled one of the largest publicly available datasets from Microsoft Academic Search, including, as of 2015, more than 5 million articles (published between 1950 and 2015) and 3 million authors in the Computer Science domain. Moreover, each paper includes additional bibliographic information such as title, unique index, author(s), affiliation(s), year of publication, publication venue, related research field(s), abstract and keyword(s) [13]. (ii) `Physics`: we collected all published articles in Physical Review (PR) journals [45] from 1950 through 2012. We only use entries which include information about index, title, name of the author(s), year of publication and references. The filtered dataset contains 425, 399 valid papers and 298, 154 authors [8]. We take `CS` as a representative of all the research areas where authors prefer conference publications; whereas `Physics` serves as a representative for journal-oriented areas (see Supplementary Text for more explanation of the datasets).

Both datasets span a temporal duration of over half a century, allowing us to investigate the success trajectory of papers and authors over a long timespan. Although each of the datasets can be viewed as a proxy for monodisciplinary research activities in science, we believe that analysis of both the datasets jointly allows us to support multidisciplinary features which characterize citation dynamics in general.

2.2 Trajectory classification

We analyze the success trajectories of entities (both papers and authors) using bibliographic indicators. The shape of the trajectories can be characterized by the following four parameters: (i) number of peaks (burstiness), (ii) time of peaks (awakening times), (iii) time elapsed before a peak (length of sleep or hibernation time) and (iv) intensity of success during the time of sleep (depth of sleep).

2.2.1 Preliminary definitions

Here, we briefly mention the definition of the parameters based on which we identify different trajectories. Most of the parameters are adopted from our previous papers [9,10].

Scientific success The count of raw citations of a research endeavor has turned out to be a simple but quite effective measure for quantifying the quality of research and is thus a universally accepted measure of research success [38,43]. Following the same line, we quantify the *success* of a paper p at year t (denoted by S_p^t) by the number of citations received by p at year t. Similarly, the *success* of an author a at year t (denoted by S_a^t) is

defined by the ratio between the number of citations received by a at t (termed as C_a^t) and the number of papers published by a until t (terms as P_a^t), i.e., $S_a^t = \frac{C_a^t}{P_a^t}$.

Success trajectory A temporal trajectory of success can be represented by a set of success values ordered by time. Therefore, the success trajectory of a paper p is defined as $\mathbb{ST}(p) = \{S_p^t\}_{t=t_{\text{start}}^p}^T$, where t_{start}^p is the year of publication of the paper p and T is the current time. Similarly, the success trajectory of an author a is defined as $\mathbb{ST}(a) = \{S_a^t\}_{t=t_{\text{start}}^a}^T$, where t_{start}^a is the $6th$ year after author a wrote her first paper (according to our dataset). An initial 5-year buffer window is provided to each author after her first appearance in the research world with the assumption that, unlike for a paper, a time frame of a few years is typically required for an author to become known in her field. In particular, we assume that "5-year" time period is usually required for an author to obtain her graduation; after that she can start her career. However, our findings remain almost same with slight variation of this time window. A schematic diagram of a success trajectory is shown in Fig. 1a.

Peak A data point in the trajectory is considered to be a peak if the following two heuristics [10] are satisfied: (i) the value of the data point should be 75% higher than the maximum value present in the trajectory; (ii) two consecutive peaks should be separated by more than 2 years. Let us assume that for an entity e, the time of occurrence of the maximum peak is $t^* = \text{argmax}_{t_{\text{start}}^e \le t \le T} S_e^t$ and the height of the maximum peak is $S_e^{t^*}$. According to our heuristics, we consider a data point at t' as a peak if (i) $S_e^{t'} \ge 0.75 S_e^{t^*}$ and (ii) $\nexists t$ s.t. $(S_e^t \ge 0.75 S_e^{t^*}) \wedge (|t - t'| \le 2)$. If the second condition is encountered, we treat it as a single peak which persists for 2 years and consider the last year (the one which is higher between t and t') as the time of peak. For example, in Fig. 1a the maximum peak occurs at $t = 20$; the data point at $t = 4$ violates the first condition and the point at $t = 18$ violates the second condition. Accordingly, those are not considered as peaks. However, $S_{t=13}^e$ satisfies both conditions and qualifies as a peak, along with the data point at $t = 20$.

Awakening time The time when a peak appears in the trajectory is called the awakening time. We define $\mathbb{AT}(e)$ as a set of increasingly ordered awakening times for an entity e, i.e., $\mathbb{AT}(e) = \{t_1, t_2, \ldots, t_n | \forall t_i : t_{i+1} \ge t_i \wedge (t_{i+1} - t_i) > 2 \wedge S_e^{t_i} \ge 0.75 S_e^{t^*}\}$. For instance, in Fig. 1a, $\mathbb{AT}(e) = \{13, 20\}$.

Awakening intensity This parameter is defined by the success value of an entity at a particular awakening time. Therefore, the awakening intensity $\mathbb{AI}_e(t)$ of an entity e at t is defined as $\mathbb{AI}_e(t) = S_e^t$, where $t \in \mathbb{AT}(e)$. In Fig. 1a, $\mathbb{AI}_e(t = 13) = 27$ and $\mathbb{AI}_e(t = 20) = 36$.

Hibernation time The time elapsed between two consecutive awakening times is called the hibernation time or the length of sleep. It also includes the time gap between t_{start}^e and the first awakening time. So hibernation time of an entity e is denoted by $\mathbb{HT}(e) = \{(t_1 - t_{\text{start}}^e), (t_2 - t_1), \ldots, (t_n - t_{n-1})\}$, where $t_i \in \{\mathbb{AT}(e) \cup t_{\text{start}}^e\}$. In Fig. 1a, $\mathbb{HT}(e) = \{13, 7\}$.

Depth of a sleep The depth of a sleep is defined by the average number of citations received per year by an entity during the corresponding hibernation time (excluding the awakening time). It is denoted as $\mathbb{DS}_e(t_i, t_j) = \frac{1}{t_j - t_i} \sum_{t=t_i}^{t_j - 1} S_e^t$, where $(t_j \ge t_i)$ and $(t_i, t_j \in \{\mathbb{AT}(e) \cup t_{\text{start}}^e\})$. In Fig. 1a, $\mathbb{DS}_e(t = 0, t = 13) = 7.1$ (approx.).

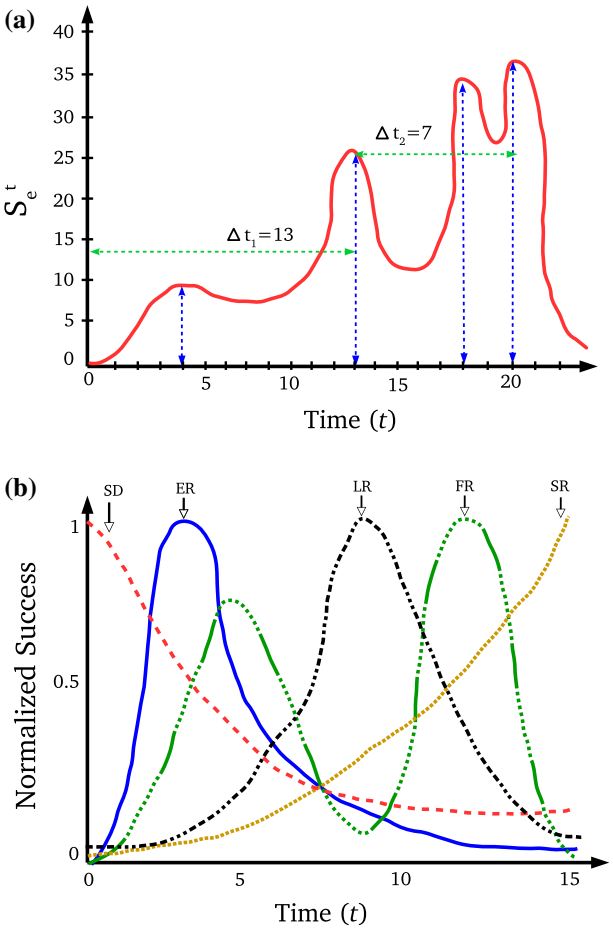

Fig. 1 **a** A schematic diagram of a success trajectory. There are two peaks (at $t = 13$ and $t = 20$) based on our heuristics of peak detection. This trajectory belongs to the "Frequent Riser" (FR) category. **b** Schematic examples demonstrating different patterns of success trajectories. We exclude Oth category because it does not follow any particular pattern

2.2.2 Heuristics and six success trajectories

We attempt to classify the trajectory $\mathbb{ST}(e)$ of an entity e (which is characterized by t_{start}, $\mathbb{AT}$, $\mathbb{AI}$, $\mathbb{HT}$, and $\mathbb{DS}$) into one of the following six categories (see Fig. 1b):

(i) *Early Riser* (ER) The trajectory of an entity should contain *only one peak* and the peak should appear within the *first 5 years* (including the fifth year) of its appearance (but not at the first year), i.e., $|\mathbb{AT}(e)| = 1$ and $t \in \mathbb{AT}(e) : (t \neq t^e_{\text{start}}) \wedge (t - t^e_{\text{start}}) \leq 5$. Moreover, the entity should receive at least one citation (on average) per year throughout its lifetime, i.e., the following condition is satisfied:

$$\left(\frac{1}{T - t^e_{\text{start}}} \sum_{t=t^e_{\text{start}}}^{T} S^t_e\right) \geq 1 \qquad \text{(We call it } \ell\text{-condition)} \tag{1}$$

(ii) *Late Riser* (LR) The trajectory of an entity should satisfy the ℓ-condition (mentioned in Equation 1) and should contain *only one peak*. The peak itself should occur 5 years (excluding the fifth year) after it appeared (but not at the last year when the observation is conducted), i.e., $|\mathbb{AT}(e)| = 1$ and $t \in \mathbb{AT}(e) : t \neq T \wedge (t - t^e_{\text{start}}) > 5$.

(iii) *Frequent Riser* (FR) The trajectory of an entity should satisfy the ℓ-condition and should also contain *multiple peaks* at different time points of the life span, i.e., $|\mathbb{AT}(e)| \geq 2$.

(iv) *Steady Riser* (SR) The trajectory of an entity should satisfy the ℓ-condition and should also contain *only one peak*. The peak should occur at the time of observation T, i.e., $|\mathbb{AT}(e)| = 1$ and $t \in \mathbb{AT}(e) : t = T$. Moreover, it should accumulate a sufficient amount of citations before the occurrence of peak ($\mathbb{DS}_e(t^e_{\text{start}}, T) \geq \delta$, where δ is a threshold parameter, and empirically set as $\frac{0.3}{(T - t_{\text{start}})} S^T_e$). Ideally, the trajectory should increase monotonically from the beginning until the time of observation.

(v) *Steady Dropper* (SD) The trajectory of an entity should satisfy the ℓ-condition and should contain *only one peak*. The peak should occur at the first year, i.e., $|\mathbb{AT}(e)| = 1$ and $t \in \mathbb{AT}(e) : t = t^e_{\text{start}}$. Moreover, it should also accumulate a sufficient amount of citations after the occurrence of the peak as mentioned for SR. Ideally, the trajectory should follow a monotonically decreasing pattern.

(vi) *Other* (Oth) The trajectory of an entity does not satisfy the ℓ-condition. We set this as the minimum condition for an entity to be qualified into one of the above five categories. An entity belonging to this category implies a lack of statistically significant evidence in support of classifying it in any of the previous categories. We therefore opt to classify such entities into an entirely separate group.

We consider only trajectories which have at least 10 years of citation history at the year of observation ($\forall e : T - t^e_{\text{start}} \geq 10$). Given a $\mathbb{ST}(e)$, we scale the data points by normalizing them with the maximum value present in it (i.e., $\hat{\mathbb{ST}}_i(e) = \frac{\mathbb{ST}_i(e)}{\max_{j=t^e_{\text{start}}}^{T} \mathbb{ST}_j(e)}$) so that the awakening intensity of the maximum peak becomes one for all the entities, and thus we obtain a normalized success trajectory $\hat{\mathbb{ST}}(e)$. Finally, we run a local-peak-detection algorithm to detect peaks in the normalized trajectory by following the rules mentioned earlier for different categories and map each trajectory into any one of the categories.

2.3 Measure of interdisciplinarity ($\mathbb{H}$)

Chakraborty et al. [11] propose three entropy-based measures to compute interdisciplinarity. Following this, we measure interdisciplinarity of a number of papers, say P, from the information of their research fields. In the CS dataset, there are 24 fields and each paper is labeled with field information (see Table S2 in Supplementary Text). Let us assume that the fields are $\{F_i\}_{i=1}^{24}$. The entropy $\mathbb{H}$ of P is measured as follows: $\mathbb{H} = -\sum_{i=1}^{24} \frac{C_i}{|P|} \log \frac{C_i}{|P|} . \delta(F_i, P)$, where C_i is the count of papers in P tagged with field F_i, and $\delta(F_i, P)$ is the Kronecker delta (returns 1 if there is at least one paper of field F_i in P, 0 otherwise).

3 Results

In this section, we present the fundamental results of our study.

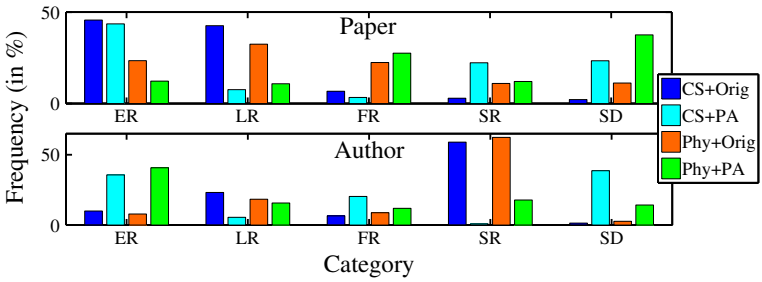

Fig. 2 Population (in %) of each category for papers and authors in the original citation dataset ("Orig") and the one generated by the Preferential Attachment model ("PA"). Here, we remove all the entities in the OT category. The proportions of entities in OT category are as follows: Paper (CS: Orig: 46.23, PA: 54.33; Physics: Orig: 41.98, PA: 55.21) and Author (CS: Orig: 34.12, PA: 44.76; Physics: Orig: 39.21, PA: 51.24)

Table 1 Characterizing entities of different categories for CS (all entries before ;) and Physics (all entries after ;) datasets. Since the Physics dataset does not have any conference information, all corresponding fields are marked as N.A. For each category, we calculate the percentage of the conference papers out of all the papers in that category

Statistics	ER	LR	FR	SR	SD
Paper					
% of conf. papers	64.35; N.A.	39.03; N.A.	39.89; N.A.	60.73; N.A.	65.26; N.A.
Avg. no. of citations	8.34; 10.56	13.45; 16.75	12.45; 14.37	56.43; 76.34	2.23; 2.43
Author					
% of conf. papers	68.36; N.A.	43.22; N.A.	51.98; N.A.	39.08; N.A.	76.09; N.A.
Avg. h-index	4.69; 3.87	4.15; 4.49	4.86; 4.21	6.10; 6.36	2.93; 3.01
Avg. age	13.24; 12.90	19.13; 16.54	20.34; 19.87	21.10; 23.43	16.69; 17.86
% of self-citations	31.01; 34.58	30.30; 28.64	25.71; 25.12	26.14; 25.65	32.67; 36.54

3.1 Category frequency

We study the population of entities in each of these categories for two datasets. Figure 2 shows that if the OT category (which is the numerical majority) is discarded from the analysis, early-rising papers are the most prevalent in CS. Further investigation suggests that most of the early-rising papers (64.35%) are conference papers (Table 1) and vice versa (i.e., 56% conference papers are early-rising papers). This population might be domain specific—for instance, ER papers dominate in Computer Science which is mostly a conference-oriented field of research. This result corroborates with the earlier claim [12] that papers published in conferences often get quick visibility after appearance and thus are cited most rapidly in the initial time period. However, after a few years their importance drastically diminishes and relevant citations tend to disappear. As expected, late-rising papers are mostly prevalent in Physics; at the same time late-rising and frequent-rising papers are mostly journal papers (Table 1). A similar reason can explain this phenomenon—journal papers, due to the lack of quick visibility that conference papers get, often take some time to get noticed by researchers and thus start receiving citations many years after publication.

One might relate the late-rising papers with the phenomenon called "sleeping beauties" in science [31]. To test how efficient our heuristics are in identifying sleeping beauties, we

consider 12 revived classic papers (published in Physical Review journals) identified by Redner [45] as a gold-standard list. These papers are well-cited old papers with the bulk of their citations occurring long after their publication. We take the papers categorized as LR in the Physics dataset, rank them based on citations and consider the top 12 papers. Remarkably, our method is able to identify 8 papers out of 12 revived papers, with the top two papers being the exact same ones observed by Redner. This essentially shows the effectiveness of our manually curated rules for classifying trajectories.

Figure 2 also depicts that most of the authors tend to be Steady Risers which seems to fit well with the intuition that, in general, it takes time for researchers to gradually gain experience and exposure that in turn leads to consistent growth in their success. However, proper analysis of the other three categories, namely ER, LR and FR which have significant presence in our datasets, also unfolds some interesting patterns. First, close observation reveals that no author category is biased toward any particular paper category (see Table S4 in Supplementary Text). We may then notice that the research age (defined by the time difference between publication of last paper minus first paper) of ER tends to be smaller when compared to the other categories, which in turn may result in citation count decay (Table 1). Similarly, for LR we observe that the time gap between the last publication and the occurrence of a peak is significantly smaller (CS: 2.34 years, Physics: 3.43 years on average) than others (CS: 6.14 years, Physics: 8.23 years on average). Both these results corroborate the fact that these authors might stop publishing papers after a certain time period, which results in the sharp decay. We shall analyze the possible reasons behind the trajectory fall later in this article. The characteristics of FR category will also be analyzed separately later.

Findings 1

- Conference papers in CS receive maximum citations early after publication.
- Journal papers take some time to get sufficient visibility.
- Population in each category is domain-dependent.
- Most of the researchers are Steady Risers.

3.2 Statistical significance of our categorization

One might assume from the results of the earlier section that the phenomenon of universal trajectories could in principle be described by a simple statistical mechanism such as "Preferential Attachment" [1]. To address this point, we build the citation network based on the Preferential Attachment model (PA) [1] as it is one of the most fundamental ideas used to model the citation histories of papers.

Preferential Attachment model [1] starts with considering an original network obtained from the initial 5 years in the dataset. For each following year t until the end, n_t papers are added, and each paper p brings r_p references. n_t is set to the number of papers in the dataset actually published in year t, and r_p corresponds to the number of references of one of the papers in that set. As the papers are progressively added to the citation network, the references they contain are linked to previously published papers chosen with probability proportional to one plus the number of citations those papers already have.

In Fig. 2, we observe that the PA model fails to bring out the proportion of each category. Sometimes, the PA model overestimates certain categories (such as for both papers and authors, the bar corresponding to SD category obtained from PA model is longer than that from the original network); while for a few categories it heavily underestimates the population (such as LR for papers, SR for authors). These results conform to the observation

that the trajectory categories are not the results of any controlled mechanism, rather they are produced via an inherent complex dynamics which control the citation distribution over the years.

Findings 2

- The traditional growth model fails to mimic these categorizations, most notably the behavior of "Late Riser" and "Steady Riser".

3.3 Perceived importance of each category

The acceptance of a research endeavor is largely determined by the raw citation count—the more an article receives citations from other papers, the larger its perceived utility for the community. Similarly, researchers are often judged by several bibliographic indicators, such as the h-index [30], the i-10 index, and the g-index [18].

In this section, we ask the question of which category is most admired overall in terms of citation count. In Table 1, we notice that, on average, steady-rising papers receive a large number of citations, which follow common intuition. However, the second most "prestigious" category seems to be LR. This sharply contradicts the perceived importance of "first-mover advantage", which is based on the idea that either papers start to accumulate citations in the early stages of their lifetime or never accumulate a significant number of citations [39]. In fact, our result implies that if a paper somehow manages to attract people later on in its lifetime, it either receives higher appreciation (compared to the ones admired earlier on) and then gradually falls (resulting in a LR) or it keeps getting citations at the same pace (resulting in a SR). However, if the citation count starts diminishing, it rarely recovers (resulting in a SD).

Table 1 shows a similar trend for authors. Steady Risers achieve highest h-index in general. In CS, SR is followed by FR; the reason might be the cumulative effect of many early-rising conference papers which result in high h-index (see Table S4 in Supplementary Text). However, due to the plethora of journal papers, the Late Risers in Physics achieve a higher h-index compared to FR. Interestingly, according to our analysis, "early rising" does not always determine whether a researcher will become popular in terms of the h-index. In fact, our results mostly point toward the fact that "consistent growth" is always important in order to maintain a stable position in research.

Another major issue that people are concerned with is self-citation [19,34]. It can falsely escalate the perception of an article's or a researcher's scientific impact, particularly when the article has been written by many authors, thus artificially increasing the overall number of citations [16]. There have been calls to remove self-citations from citation-rate calculations [47]. Here, we separate ourselves from this debate. Rather, we show which category is the most resilient to self-citations. In Table 2 we observe that frequent-rising papers suffer less from self-citation, and the second least affected category is late-rising papers. Steady-rising and steady-dropping papers appear to be more affected by it. The reason behind this could be that these papers initially receive a lot of self-citations in order to get noticed by others (see Fig. 3); some of them earn appreciation later, with citations coming from other papers (resulting in a SR), while others might not receive the expected attention and gradually lose their importance (resulting in a SD).

When it comes to authors, we observe that the Steady Riser is the most stable category without self-citations, immediately followed by the Late Riser. The steady droppers, exhibiting a large number of self-citations, would have moved to the OT category in 19.5% of cases

Table 2 Confusion matrices indicating the transition of population (in %) from one category to another after removing self-citations for CS (before) and Physics (after)

Category	ER	LR	FR	SR	SD	OT
Paper						
ER	0.72; 0.82	0.10; 0.01	0.03; 0.04	0.01; 0.00	0.00; 0.00	0.15; 0.09
LR	0.02; 0.00	**0.81; 0.85**	0.04; 0.05	0.00; 0.00	0.10; 0.09	0.11; 0.01
FR	0.01; 0.01	0.06; 0.03	**0.86; 0.89**	0.00; 0.00	0.01; 0.05	0.06; 0.30
SR	0.05; 0.04	0.04; 0.06	0.00; 0.00	0.71; 0.75	0.00; 0.00	0.20; 0.19
SD	0.00; 0.00	0.05; 0.07	0.05; 0.06	0.05; 0.05	0.67; 0.61	**0.18; 0.21**
			Author			
ER	0.81; 0.84	0.01; 0.01	0.02; 0.03	0.05; 0.05	0.01; 0.03	0.10; 0.04
LR	0.01; 0.02	**0.89; 0.86**	0.01; 0.02	0.04; 0.05	0.01; 0.01	0.04; 0.05
FR	0.04; 0.01	0.01; 0.02	0.84; 0.85	0.01; 0.01	0.00; 0.01	0.10; 0.07
SR	0.00; 0.03	0.00; 0.01	0.02; 0.03	**0.93; 0.90**	0.01; 0.02	0.04; 0.01
SD	0.01; 0.00	0.01; 0.02	0.00; 0.01	0.04; 0.02	0.74; 0.76	**0.20; 0.19**

The row indicates the actual category and the column indicates the category after removing self-citations. Important results are highlighted in bold text. Note that there is no row for OT since entities in this category can never move to other categories due to the deletion of citations

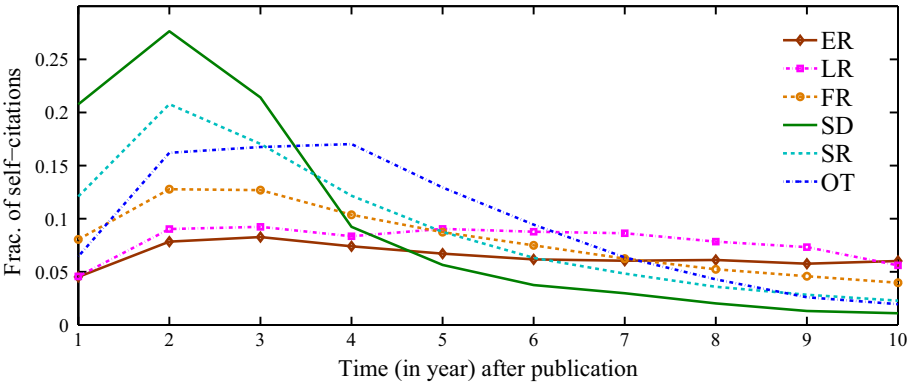

Fig. 3 Fraction of self-citations per paper in different categories over 10 years after publication

if self-citations were removed. Therefore, one might conclude that steady droppers persist because of their own citations.

Findings 3

- Steady Risers tend to receive the highest number of citations, emphasizing the fact that the "consistency" in research is most important.
- Late-rising papers stand at second, which leads to a sharp contradiction with the "first-mover advantage" phenomenon.
- Authors may avoid earlier rising tendency in order to attain success in future.
- In research, self-citations are self-promotion, which either result in high visibility of research or might diminish the popularity.

3.4 Category stability

The behavior of incoming citations of an entity over time indicates whether the entity stands the test of time. Moreover, the temporal dynamics of citations controls the shape of the trajectory of each entity. In this section, we pose the question how stable these categories are over time. More specifically, if an entity is marked with a certain category, would it ever change over time? A long citation history corresponding to each entity allows us to conduct a broad study across different time stamps. For each entity, we assign it into one of the categories at $T + 10$, $T + 15$ and $T + 20$ (where $T = t_{\text{start}}$ for the entity). Note that when measuring the success of an entity at a certain time t, we only consider the history on or prior to t. The alluvial diagram of Fig. 4 shows the stability of the categories for CS (the results are almost identical to that of Physics; see Supplementary Text).

For papers Fig. 4a, we observe that all categories are almost stable. SD seems to be the most stable category among all, i.e., it is less likely that the mass in this category would get fragmented over time, hinting to the fact that the impact of a paper once starts decreasing is rarely recovered. The proportion in FR category seems to be increasing with time, with a significant mass of OT migrating toward FR. At the same time, papers in OT also turn out to be SR over time due to the manifestation of peaks in the later time period. This again confirms our earlier observation in the earlier section (and Table 1) and contradicts the idea of "first-mover advantage".

The migration is quite prominent (the categories are less stable) for authors as shown in Fig. 4b. Initially, the proportion seems to be almost equal for all the categories. However, with time SR and OT start attracting major mass from SD, LR and FR. The volume of LR and SD tends to be diminishing over time. The drifting of ER and SD to OT might be explained by the self-citation phenomenon—although due to the self-citations ER and SD manage to maintain their existence in the initial years, they fail to keep their presence later and gradually migrate to OT. The migration from ER to SR seems to be quite persistent over time; one of the reasons could be due to the career switch from "student" to "mentor". In 20% of cases, LR retains the success after the peak and moves to SR over time. This analysis indeed demonstrates a systematic approach that examines the transition from one category to another when an increase of citations takes place.

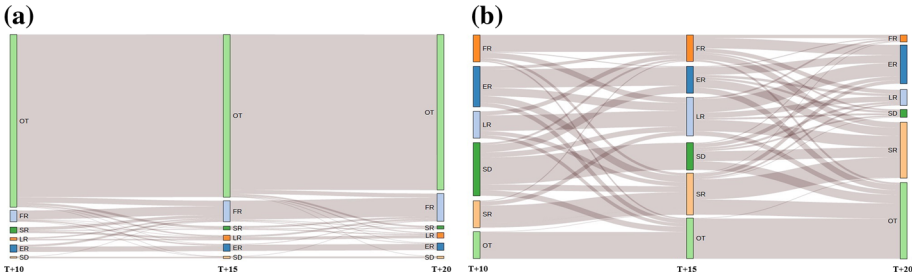

Fig. 4 Alluvial diagram showing the stability of different categories for **a** papers and **b** authors in CS, i.e., how entities from one category migrate to another category after 15 and 20 years (results are identical for Physics and reported in Supplementary Text). The colored blocks correspond to different categories. *Block size* indicates the number of entities in a category, and the *shaded waves* joining the regions represent flow of entities between the regions, such that the width of the flow corresponds to the fraction of papers. The total width of incoming flows is equal to the width of the corresponding region. The actual values used to draw this diagram are presented in the Supplementary Text

Findings 4

- A paper once categorized as steady dropper is less likely to be shifted to another category.
- There is enough evidence of less-cited papers becoming popular later in time.
- Generally, Steady Risers keep the signature of "consistency" early in their career.
- It seems that with time "Steady Risers" and "others" categories dominate the entire population.

3.5 Triggers of trajectory awakening

To find what causes the awakening of a trajectory, we need to perform a case-by-case analysis of all the trajectories. In particular, we analyze the source and the destination of citations emitted at the awakening times.

In Fig. 5a, we plot the awakening time distribution for papers in the ER, LR and FR categories. For ER, in most of the cases the peak occurs at the third year after publication, while for LR it occurs somewhere between the sixth and seventh years. For the FR category, we find two peaks on average, where the first peak typically occurs at the fourth year, while the second peak occurs at the eighth year. Figure 5b shows that unlike CS, there are significant number of FR papers in Physics with more than two peaks (we shall discuss the characteristics of FR later in more details). The immediate question would then be: what is the hidden factor which makes a paper move from its hibernation state to an active state, such that it starts receiving significant attention so many years after publication? The field information associated with each paper in CS allows us to answer this question. To that end, we concentrate on the LR papers and measure the *interdisciplinarity* ($\mathbb{H}$) [11] (see Materials and Methods) of those

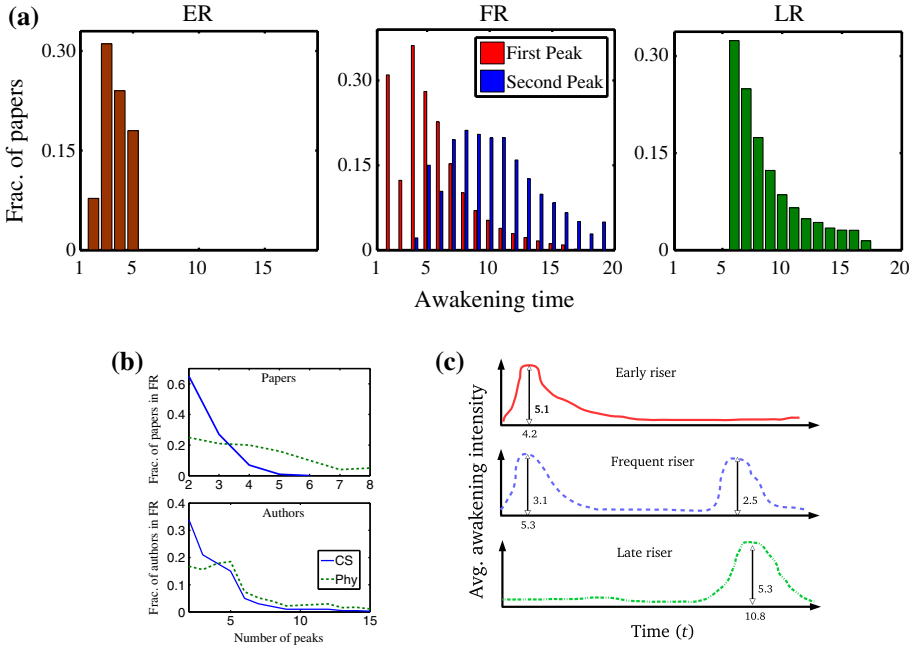

Fig. 5 **a** Awakening time and **b** peak distribution of entities in different categories. **c** Hypothetical lines showing number of peaks, average awakening intensity and average awakening time for ER, LR and FR categories in CS. FR category seems to behave as an intermediary between ER and LR categories

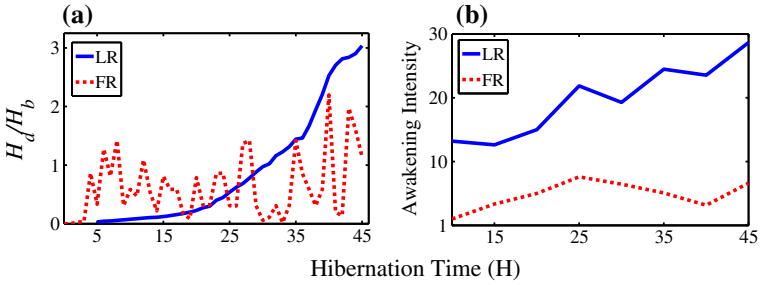

Fig. 6 **a** Ratio of the interdisciplinarity of citing papers during and before the awakening time, and **b** average awakening intensity for LR and FR categories for different values of awakening time. For LR, the ratio and the awakening intensity increases with the awakening time; however, there is no such correlation for the FR category (see Supplementary Text for the individual data points)

papers citing each LR paper *before* the awakening time ($\mathbb{H}_b$) and *during* the awakening time ($\mathbb{H}_d$). In Fig. 6a, we plot the ratio of $\mathbb{H}_d$ and $\mathbb{H}_b$ with respect to the hibernation time. The more the value of the ratio, the stronger the claim that the papers citing a LR paper at its awakening time are more interdisciplinary than the same before that time. Surprisingly, we observe that the longer the hibernation time of LR, the more citations the paper gets from diverse research fields. Therefore, the sudden awakening of a LR paper may be explained by the fact that the paper under consideration is suddenly "discovered" as being relevant by other research communities. The following case study might strengthen our claim: we consider an extreme case—a paper written by Garfield, titled "Citation indices for science; a new dimension in documentation through association of ideas" [20] in 1955. After almost 50 years in hibernation state, it suddenly became noticed by the community in 1999 due to the citation of the famous paper of Kleinberg "Authoritative Sources in a Hyperlinked Environment" [33]. Afterward, it received an enormous amount of citations from papers related to journal impact factor and similar subjects. Interestingly, there is a positive correlation between hibernation time and awakening intensity, i.e., the more a paper remains in the hibernation state, the more it becomes popular once awake (Fig. 6b). Both these results might suggest that a "premature" paper may not be able to attract enough attention from the research community when published; however, when people understand its importance, they appreciate it much more than usual. A current example of this trend is Einstein's research on the existence of gravitational waves, which occurred around 100 years ago yet was only recently confirmed and heavily cited. It should be noted, however, that the dynamics are slightly different for FR papers and will be discussed later in this paper.

On the other hand, we speculate that the peak in the trajectory of an author (particularly for LR and FR) occurs due to a sudden popularity of one or more papers by the author at the awakening time, instead of the commonly held notion that this occurs because of the cumulative popularity of many of her papers. To verify this, we consider all citations at each awakening time for LR and FR and measure the entropy of the citing papers. For example, let us assume that there are 10 citations at a particular awakening time, pointing to two papers in two different ways: (i) 5 citations each for 2 papers, (ii) 8 citations for one paper and 2 for another paper. The entropy in the first case is higher than in the second case. Note that the second case represents our speculation. In Fig. 7 we plot the entropy value at each awakening time and compare it with the case where the citations are uniformly distributed among the cited papers. The low entropy value corroborates our speculation that the formation of the peak is a result of few papers which receive a large amount of citations at the awakening

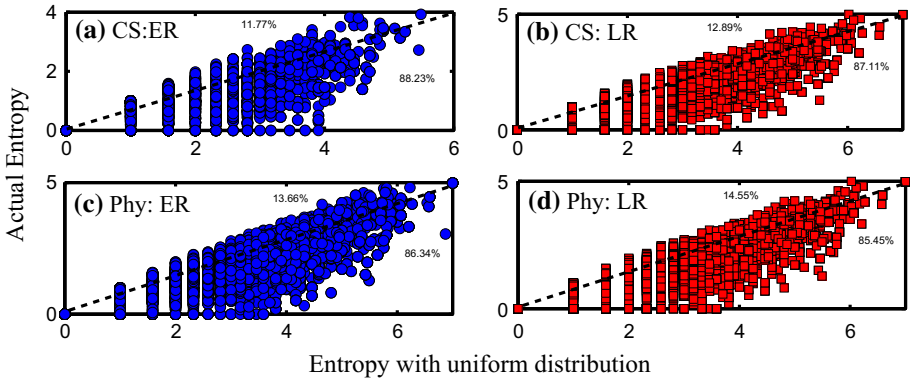

Fig. 7 Entropy of the citation distribution at the time of peak. We show the results of ER and LR for two datasets. On the x-axis, we plot the entropy of the citations if those pointed uniformly to other papers. On the y-axis, we plot the actual entropy values. The actual entropy is much smaller than the uniform case, indicating that only a few papers are responsible for the creation of peaks. The percentage of points above and below the diagonal line is shown in all the subfigures

time. For example, one of the highest cited authors in the LR category is Keir Fraser, who received 80% of the citations in 2010 for his paper "Xen and the art of virtualization" [2].

Findings 5

- Sleeping beauties are triggered by the citations from diverse research communities and may be responsible for the birth of a new research area.
- Researchers' profiles exhibit sudden peaks in different times owing to a small number of highly cited papers instead of the cumulative citations of many papers.

3.6 Triggers of trajectory decline

The question of why a paper's trajectory decays has for a long time been answered by the claim that the paper does not have enough quality to sustain itself in the research world. However, for the authors' case, finding appropriate reasons for the decline might not be so straightforward. In this section, we analyze three possible reasons that attempt to explain the decline of success of authors (note that these reasons may not be mutually exclusive):

(i) Collaboration In the current era of multidisciplinary research, researchers are keen to collaborate with others. To analyze the effect of collaboration, we identify, for each author, her most prominent collaborator (in terms of h-index) before the decline. We observe that during the time of decay, 46% (37%) of authors are unable to retain their most prominent collaborators in CS (Physics). We hypothesize that this is one of the reasons for the decay of success.

(ii) Retirement In Table 1, we observed that FR and SR are involved in research for longer time frames when compared to the others. Here, we intend to measure the time gap between the retirement (the time at which the author wrote her last paper) and the last awakening time of each author. If the time is too short, it might indicate that the reason for the decline of success is due to the fact that the author is no longer involved with research. We observe that in 64% (56%) of cases in CS (Physics), the authors retire within 5 years of the last awakening time and the rate of publications significantly diminishes after

retirement compared to the same rate before the awakening time. Therefore, retirement could be another reason for the decline of trajectory.

(iii) Trade-off between quality and quantity This reason stems from today's style of research—"publish or perish" [3]—the pressure in Academia to rapidly and continuously publish academic work to sustain oneself in the research world. We are curious to see how the peer pressure of "quantity" abolishes the "quality" of research. We find that for around 82% (79%) of cases, the value of success for ER, LR and SD drops due to the enormous volume of individual publications, which result in overshadowing the effect of incoming citations for CS (Physics). Interestingly, for both datasets (CS; Physics) the rate of publications of SR (2.06; 1.27) is the smallest among others (on average 4.32; 3.29), which indicates that SR tends to emphasize *quality*, rather than quantity. Frequent Risers seem to prefer producing quality papers with gaps of 3–5 years, which leads to sudden peaks in their success trajectories.

Findings 6

- Collaborative research might be a good way for sustaining oneself over a longer period of time.
- One should maintain a balance between quality and quantity of research in order to maintain steady success.

3.7 More about the Frequent Riser category

Entities falling in the FR category exhibit more than one peak at different time points in the trajectory, which significantly differentiates this category from all others. In order to capture the average behavior, we first plot the distribution of the number of peaks for FR papers in Fig. 5b. We observe that the maximum number of peaks is 6 (8) for CS (Physics). Moreover, the distribution of peaks in Physics seems to be quite uniform compared to that of CS. The possible reason could be the effect of majority of the journal papers in Physics, which unlike the conference papers generally sustain for longer time span that in turn helps these papers getting attentions in a regular basis. In Fig. 5c, we show the average awakening intensity and the average time of occurrences of the peaks for ER, FR and LR categories. A deeper analysis unfolds three interesting observations—(i) most of the papers (65.04%) in FR category have two peaks on an average in the timeline of the success trajectory; (ii) the sum of the average awakening intensity of first two peaks in FR category (i.e., $3.1 + 2.5 = 5.6$) is (nearly) similar to the awakening intensity of the peak for ER and LR categories (5.1 and 5.3, respectively); (iii) the average difference between the time of occurrences of the first two peaks in FR category (i.e., $12.1 - 5.3 = 6.8$) is (nearly) similar to the difference of the occurrence of the peak in LR and ER categories (i.e., $10.8 - 4.2 = 6.6$). From these observations, one could argue that FR behaves as an intermediary between ER and LR categories.

For further analysis, we start looking at the co-citation [5,49] (see the definition in Supplementary Text) pattern between the FR paper and the other papers. Specifically, for a particular FR paper X, we retrieve the paper Y that (a) cites X at its awakening time, and (b) is such that both X and Y are cited together heavily after the citation of Y to X. This might indicate that paper Y is responsible for pulling out X from its hibernation state. The papers which wake up a "sleeping beauty" are often known as "princes of the sleeping beauty" [36]. We intend to identify all such papers Y for different awakening times. To demonstrate this, we choose a classic example—the most cited paper "Gradient based learning applied to document recognition" [35] in the FR category, where the concept of "Deep Learning" was

introduced for the first time. Since the year of publication (1998), it did not receive sufficient citations, until a citation of the famous book written by Schölkopf and Smola [46] in 2001, where the authors explained a detailed way of understanding Support Vector Machines and Kernel tricks. Again, after 5 years, it received a citation from another popular paper written by Hinton et al. [29], and both these papers jointly started getting a lot of citations. This example again emphasizes the fact that "premature" papers tend to not get applause during the publication time, yet if they really have potential to explore something novel, they will probably get accepted with acclamation in the future. But the fact which makes the FR papers different from LR papers is that the citations of the former category of papers are mostly confined within a particular domain and rarely get citations from different domains, whereas the papers in the latter category are mostly responsible for "ground-breaking" research which might trigger a completely new research area in the future (Fig. 6).

Findings 7

- Frequent Risers seem to exhibit an intermediary behavior between early risers and Late Risers.
- The frequency of awakening of papers is quite prominent in the journal-oriented domains such as Physics.
- Few highly cited papers are responsible for pulling a certain paper out from its hibernation state, which in turn receive enormous attention together.

3.8 Can machine learning categorize citation trajectories?

All the discoveries presented in this paper so far are based on our hand-crafted criteria of categorizing a trajectory into one of the six classes as mentioned in Sect. 2.2; although it is worth mentioning that most of the criteria are soft criteria [9,10], and the results do not vary much if the thresholds are slightly modified. However, one might be interested to devise an automated technique to categorize a citation trajectory. Note that since the time span for all the trajectory is not same (as the year of publication varies across papers), we can not use any existing time series data clustering framework [44] to categorize the trajectories. Therefore, we conduct a machine learning-based approach to categorize the trajectories automatically.

Due to the huge number of citation trajectories present in our datasets, here we present a small-scale analysis—we randomly selected 500 paper trajectories and 500 author trajectories each from CS and Physics datasets, constituting 2000 trajectories in total. We made sure that the selected trajectories cover all types of trajectories (based of hand-crafted criteria). Three human annotators[1] were shown the visualization of each trajectory and told to annotate each trajectory into one of the six categories. Note that we did not provide the annotators our definition of the categories mentioned in Sect. 2.2.2. Rather, we only provided them the following definitions of the categories:

- *Early Riser* Trajectories following this pattern show an "early peak" after their publication year.
- *Late Riser* Trajectories following this pattern show a "late peak" much later after their publication year.
- *Frequent Riser* Trajectories following this pattern may contain "multiple peaks" in different time points of their entire span.
- *Steady Riser* Trajectories following this pattern exhibit a "consistent growth" throughout their entire time span.

[1] The annotators are experts on Bibliographic search.

- *Steady Dropper* Trajectories following this pattern exhibit a "gradual decay" as the time progresses.
- *Other* Trajectories which do not fall in any of the previous five categories are assigned to this category.

Note that the terms such as "early peak", "late peak", 'multiple peaks", "consistent growth" and "gradual decay" are subjective and depend upon annotators' perception. We intended to investigate how the categorization based on our hand-crafted rules are correlated with the output generated from different machine learning models trained on human annotations.

Out of total 2000 annotations, 1700 (`CS`: Paper: 428, Author: 429; `Physics`: Paper: 427, Author: 416) are such annotated trajectories where at least two annotators agreed with the categories. We also observed that "other" category has the most disagreement. The (average) inter annotators' agreement is 0.83 based on Cohen's kappa, which might be regarded to be a high agreement [17]. In supplementary, we present confusion matrices to show the agreement between human annotation and rule-based annotation for 1700 trajectories.

Table 3 Sets of features used in paper (P) and author (A) trajectory classification models

	Feature	Description	Model
Trajectory-based	# of peaks	No. of peaks based on our heuristics	P&A
	First awakening time	Time difference between the first peak and the starting year	P&A
	Last awakening time	Time difference between the first peak and the last year	P&A
	Avg. hibernation time	See the definition in Sect. 2.2.2	P&A
	Avg. depth of sleep	See the definition in Sect. 2.2.2	P&A
	Max. depth of sleep	See the definition in Sect. 2.2.2	P&A
	Avg. awakening intensity	See the definition in Sect. 2.2.2	P&A
	Max. awakening intensity	See the definition in Sect. 2.2.2	P&A
	FirstPeak_in5	Does the first peak appear within first 5 years of the staring time	P&A
	AvgCite	Avg. number of citations per year	P&A
Author-based	Efficiency	# of citations per author	P&A
	h-index	Avg. *h*-index per author	P&A
	Sociality	Avg. # of coauthors per author	P&A
	Productive	# of papers per author	P&A
	Frac. of conf. papers	Frac. of conference papers per author	A
	Frac. of journal papers	Frac. of journal papers per author	A
Paper-based	team size	# of authors of the paper	P
	IF	Impact factor of the venue (journals/conferences)	P
	Is_conf	Is the paper published in a conference?	P
	RefCount	# of references of the paper	P
	# of citations	# of citations received by the paper so far	P

Table 4 Average F-score of three classifiers after tenfold cross validation

	ER	LR	FR	SR	SD	OT	Overall
SVM							
CS							
Paper	0.86	0.84	0.78	0.87	0.78	0.62	0.79
Author	0.83	0.80	0.76	0.85	0.81	0.69	0.79
Physics							
Paper	0.89	0.88	0.77	0.83	0.87	0.73	0.79
Author	0.85	0.88	0.82	0.85	0.81	0.76	0.84
Decision Tree							
CS							
Paper	0.79	0.86	0.73	0.71	0.76	0.60	0.74
Author	0.75	0.72	0.71	0.78	0.73	0.63	0.72
Physics							
Paper	0.82	0.76	0.71	0.72	0.77	0.60	0.73
Author	0.72	0.79	0.70	0.73	0.74	0.63	0.72
Logistic Regression							
CS							
Paper	0.81	0.73	0.73	0.79	0.71	0.58	0.72
Author	0.77	0.79	0.74	0.82	0.72	0.65	0.75
Physics							
Paper	0.73	0.84	0.73	0.79	0.81	0.62	0.75
Author	0.73	0.75	0.71	0.76	0.80	0.63	0.73

We report the accuracy corresponding to each category separately

We then use three standard machine learning classifiers—Support Vector Machine (SVM), Decision Tree and Logistic Regression. We perform hyper-parameter optimization in order to find the parameters that generate the best results. For instance, we use CART with Gini gain criteria for DT, multinomial logistic regression and SVM with linear kernel.

For each trajectory, we consider three sets of features mentioned in Table 3—trajectory based, author based and paper based (we also mention in Table 3 which features are used in which classification model). We used our hand-crafted heuristics to extract these features. The motivation of this study is to show how manually-designed trajectory-based metrics as features help us predict the human annotated categories. The average accuracies of the classifiers are reported after stratified tenfold cross validation.

Table 4 reports the average F-score of the classifiers for individual categories separately. Overall, SVM outperforms others with an average F-score of 0.79 for both papers and authors in CS, and 0.79 and 0.84 for papers and authors in Physics, respectively. We also observe that trajectory-based features (especially number of peaks, average and maximum awakening intensity) are more important in this categorization task (see Supplementary for the detailed feature importance). This analysis indeed demonstrates that machine learning models with hand-crafted rules and other intrinsic features can provide a domain agnostic classification of success trajectories.

4 Discussion

In this study, we systematically investigated the long-range citation history of papers and authors together, with the purpose of unfolding different patterns of success trajectories. Based on the awakening process of entities, we found at least six distinct trajectories widespread across both Computer Science and Physics domains. A series of analysis conformed to the fact that these patterns are not exceptional, rather these are extreme cases emerged from an erogenous but otherwise continuous distribution, which is impossible to be described by the traditional growth models. Some of the findings may call for further elicitation.

First, the late-rising papers seemed to be mostly journal papers; conversely for early risers and conferences. But the present formulation of Impact Factor only considers the citation history of a paper's early years (either first 2 or first 5 years after publication). By doing so, the Impact Factor essentially avoids the "golden time" of a vast majority of journal papers, which starts 5 years after publication. Therefore, *we need to rethink the justification of Impact Factor formulation*.

Second, when a new growth model is introduced nowadays, it is usually validated based on the degree distribution. If the model is able to produce the power law degree distribution, it is well accepted. However, in Fig. 2 we showed that none of the existing growth models which are primarily based on Preferential Attachment and/or aging are able to mimic these categorizations. Therefore, *we stipulate that any future proposed growth model should also be validated based on these categories in order to reflect a true sense of citation dynamics*.

Third, future success of researchers should not be predicted from their early success. *Rather one should look into how consistent their careers are initially from their appearance in the field, which might lead to a better prediction of their long-term success.* This consistency can be helpful for the major decision making process such as promotion.

Fourth, in order to build prediction models such as early prediction of citations of scientific articles, predicting high-impact and seminal papers, or recommending scientific articles for a particular researcher, *one should consider the parameters such as awakening time, awakening intensity, hibernation time, depth of sleep as mentioned in this paper, which characterize the papers' trajectories*.

Apart from the directions mentioned above, a detailed investigation is needed in order to understand the general mechanisms responsible for different awakening times of different categories. Moreover, we are interested in understanding the micro-dynamics controlling the shape of Frequent Risers. These dynamics might be very different from those of other categories. Another line of direction could be to develop and validate a bibliometric framework for identifying the "princes" who wake up the late-rising and frequent-rising papers in challenge-type scientific discoveries, so as to figure out the awakening mechanisms, and promote potentially valuable but not readily accepted innovative research. This can in turn be useful as an instructive model when studying the mechanisms of scientific information flow through citations. We would also like to design a prediction model that can predict the sleeping beauties at the early stage of their publication. We hope that this paper would motivate researchers working in other disciplines to verify the conclusions presented in this paper for their disciplines.

Acknowledgements We thank Saswata Pandit, Jason Filippou, Barbara Lewis and Fabio Pierazzi for their valuable suggestions.

References

1. Barabasi A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512
2. Barham P, Dragovic B, Fraser K, Hand S, Harris T, Ho A, Neugebauer R, Pratt I, Warfield A (2003) Xen and the art of virtualization. SIGOPS Oper Syst Rev 37(5):164–177. doi:10.1145/1165389.945462
3. Beasley CJ (2005) Publish or perish. Lead Edge 24(9):872–872
4. Bharathi DG (2013) Evaluation and ranking of researchers? Bh index. PLoS ONE 8(12):e82050. doi:10.1371/journal.pone.0082050
5. Biscaro C, Giupponi C (2014) Co-authorship and bibliographic coupling network effects on citations. PLoS ONE 9(6):1–12
6. Bollen J, Crandall DJ, Junk D, Ding Y, Börner K, Collective allocation of science funding: from funding agencies to scientific agency. arXiv:1304.1067
7. Bornmann L, Daniel H (2006) Selecting scientific excellence through committee peer review–a citation analysis of publications previously published to approval or rejection of post-doctoral research fellowship applicants. Scientometrics 68(3):427–440
8. Chakraborty T, Ganguly N, Mukherjee A (2014) Rising popularity of interdisciplinary research—an analysis of citation networks. In: Sixth international conference on communication systems and networks, COMSNETS 2014, Bangalore, 6–10 Jan 2014, pp 1–6. doi:10.1109/COMSNETS.2014.6734940
9. Chakraborty T, Kumar S, Goyal P, Ganguly N, Mukherjee A (2014) Towards a stratified learning approach to predict future citation counts. In: JCDL, IEEE Computer Society, pp 351–360. http://dblp.uni-trier.de/db/conf/jcdl/jcdl2014.html#0002KGGM14
10. Chakraborty T, Kumar S, Goyal P, Ganguly N, Mukherjee A (2015) On the categorization of scientific citation profiles in computer science. Commun ACM 58(9):82–90. doi:10.1145/2701412
11. Chakraborty T, Kumar S, Reddy MD, Kumar S, Ganguly N, Mukherjee A (2013) Automatic classification and analysis of interdisciplinary fields in computer sciences. International conference on social computing (SocialCom). Alexandria, VA, pp 180–187
12. Chakraborty T, Sikdar S, Ganguly N, Mukherjee A (2014) Citation interactions among computer science fields: a quantitative route to the rise and fall of scientific research. Soc Netw Anal Min 4(1):187
13. Chakraborty T, Sikdar S, Tammana V, Ganguly N, Mukherjee A (2013) Computer science fields as ground-truth communities: their impact, rise and fall. In: Advances in social networks analysis and mining 2013, ASONAM 13, Niagara, ON, Aug 25–29, 2013, pp 426–433. doi:10.1145/2492517.2492536
14. Crespo JA, Ortuño-Ortín I, Ruiz-Castillo J (2012) The citation merit of scientific publications. PLoS ONE 7(11):1–9
15. de Solla Price D (1963) Little science, big science- and beyond (A Columbia paperback). Columbia University Press, New York
16. Della Sala S, Brooks J (2008) Multi-authors' self-citation: a further impact factor bias? Cortex 44(9):1139–45
17. Di Eugenio B, Glass M (2004) The kappa statistic: a second look. Comput Linguist 30(1):95–101. doi:10.1162/089120104773633402
18. Egghe L (2006) Theory and practise of the g-index. Scientometrics 69(1):131–152
19. Fowler J, Aksnes D (2007) Does self-citation pay? Scientometrics 72(3):427–437
20. Garfield E (1955) Citation indexes for science. A new dimension in documentation through association of ideas. Science 122: 1123–1127. http://www.garfield.library.upenn.edu/papers/science_v122v3159p108y1955.html
21. Garfield E (1980) Premature discovery or delayed recognition–why? Curr Contents 21:5–10
22. Garfield E (1989) Delayed recognition in scientific discovery: citation frequency analysis aids the search for case history. Curr nt Contents 23:3–9
23. Garfield E (1999) Journal impact factor: a brief review. CMAJ 161(8):979–980
24. Garfield E (2006) The history and meaning of the journal impact factor. JAMA 295(1):90–93
25. Gingras Y, Larivière V, Macaluso B, Robitaille J-P (2009) The effects of aging on researchers' publication and citation patterns. PLoS ONE 3(12):1–8
26. Glänzel W, Schlemmer B, Thijs B (2003) Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. Scientometrics 58(3):571–586
27. Guimera R, Uzzi B, Spiro J, Amaral L (2005) Team assembly mechanisms determine collaboration network structure and team performance. Science 308(5722):697–702
28. Hajra KB, Sen P (2005) Aging in citation networks. Phys A 346(1–2):44–48
29. Hinton GE, Osindero S, Teh Y-W (2006) A fast learning algorithm for deep belief nets. Neural Comput 18(7):1527–1554
30. Hirsch JE (2005) An index to quantify an individual's scientific research output. PNAS 102(46):16569–16572

31. Ke Q, Ferrara E, Radicchi F, Flammini A (2015) Defining and identifying sleeping beauties in science. PNAS 112(24):7426–7431
32. Kinney AL (2007) National scientific facilities and their science impact on nonbiomedical research. PNAS 104(46):17943–17947
33. Kleinberg JM (1999) Authoritative sources in a hyperlinked environment. J ACM 46(5):604–632. doi:10.1145/324133.324140
34. Kulkarni AV, Aziz B, Shams I, Busse JW (2011) Author self-citation in the general medicine literature. PLoS ONE 6(6):1–5
35. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86:2278–2324
36. Li S, Yu G, Zhang X, Zhang WF (2014) Identifying princes of sleeping beauty—knowledge mapping in discovering princes. In: International conference on management science engineering (ICMSE), Helsinki, pp 912–918
37. Liu NC, Cheng Y, Liu L (2005) Academic ranking of world universities using scientometrics–a comment to the "fatal attraction". Scientometrics 64(1):101–109
38. Meho LI (2007) The rise and rise of citation analysis. Phys World 1(20):32–36
39. Newman M (2009) The first-mover advantage in scientific publication. Europhys Lett 86:68001
40. Petersen AM, Stanley HE, Succi S (2011) Statistical regularities in the rank-citation profile of scientists. Sci Rep 1: doi:10.1038/srep00181
41. Pradhan D, Paul PS, Maheswari U, Nandi S, Chakraborty T (2016) C^3-index: revisiting author's performance measure. In: Proceedings of the 8th ACM conference on web science, WebSci 2016, Hannover, 22–25 May 2016, pp 318–319. doi:10.1145/2908131.2908185
42. Pradhan D, Paul PS, Maheswari U, Nandi S, Chakraborty T (2017) C^3-index: a pagerank based multi-faceted metric for authors' performance measurement. Scientometrics 110(1):253–273. doi:10.1007/s11192-016-2168-y
43. Radicchi F, Fortunato CS (2008) Universality of citation distributions: towards an objective measure of scientific impact. PNAS 105(45):17268–17272
44. Rani S, Sikka G (2012) Article: Recent techniques of clustering of time series data: a survey. Int J Comput Appl 52(15):1–9 full text available
45. Redner S (2005) Citation statistics from 110 years of physical review. Phys Today 58(6):49–54
46. Scholkopf B, Smola AJ (2001) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge
47. Schreiber M (2007) Self-citation corrections for the Hirsch index. Europhys Lett 78:1–6
48. Sekercioglu CH (2008) Quantifying coauthor contributions. Science 322(5900):371
49. Small H (1973) Co-citation in the scientific literature: a new measure of the relationship between two documents. J Am Soc Inf Sci 24(4):265–269. doi:10.1002/asi.4630240406
50. Sun X, Kaur J, Milojevic S, Flammini A, Menczer F (2013) Social dynamics of science. Sci Rep. doi:10.1038/srep01069
51. van Raan AFJ (2004) Sleeping beauties in science. Scientometrics 59(3):467–472
52. Wallace ML, Larivière V, Gingras Y (2012) A small world of citations? The influence of collaboration networks on citation practices. PLoS ONE 7(3):1–10
53. Wang D, Song C, Barabási A-L (2013) Quantifying long-term scientific impact. Science 342(6154):127–132
54. Wendl MC (2007) H-index: however ranked, citations need context. Nature 449(7161):403

Tanmoy Chakraborty is an Assistant Professor, Indraprastha Institute of Information Technology, Delhi (IIIT-D), India. He completed his Ph.D. as a Google India Ph.D. fellow in the Dept. of CSE, IIT Kharagpur, India, in September 2015. Afterward, he spent around one and half years at University of Maryland, College Park, USA, as a postdoctoral researcher. His Ph.D. thesis has been recognized as the best Ph.D. thesis by Xerox Research India and IBM Research India. He has also received INAE doctoral level innovation student project award, best paper runner up in ASONAM'16, best poster award in Microsoft TechVista'15. His primary research interests include Network Science, Data Mining and Data-driven cybersecurity. He has served as a PC member in various top conferences including AAAI, PAKDD, IJCAI, and reviewer of top journals including ACM TKDD, IEEE TKDE, ACM TIST, CACM. He is also co-organizer of two workshops TextGraphs-10 (NAACL'16) and SMERP (ECIR'17).

Subrata Nandi is presently working as an Associate Professor in National Institute of Technology, Durgapur, India. He is a B.Tech. in Computer Science from Calcutta University. He got his M.Tech. from Jadavpur University and Ph.D. from IIT Kharagpur. His broad interest lies in designing networking systems for developing regions. He is also interested in understanding and modeling dynamics in large-scale datasets. He published papers in PRE, AdHoc Networks, ACM Mobicom, Sigspatial, DEV, IEEE Infocom, Scientometrics, etc.