RESEARCH ARTICLE

Scientometric engineering: Exploring citation dynamics via arXiv eprints

Keisuke Okamura[1,2,3]

[1]Institute for Future Initiatives (IFI), The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[2]Ministry of Education, Culture, Sports, Science and Technology (MEXT), 3-2-2 Kasumigaseki, Chiyoda-ku, Tokyo 100-8959, Japan
[3]SciREX Center, National Graduate Institute for Policy Studies (GRIPS), 7-22-1 Roppongi, Minato-ku, Tokyo 106-8677, Japan

Keywords: arXiv, bibliometrics, citation dynamics, preprints, research evaluation, stochastic modeling

ABSTRACT

Scholarly communications have been rapidly integrated into digitized and networked open ecosystems, where preprint servers have played a pivotal role in accelerating the knowledge transfer processes. However, quantitative evidence is scarce regarding how this paradigm shift beyond the traditional journal publication system has affected the dynamics of collective attention on science. To address this issue, we investigate the citation data of more than 1.5 million eprints on arXiv (https://arxiv.org) and analyze the long-term citation trend for each discipline involved. We find that the typical growth and obsolescence patterns vary across disciplines, reflecting different publication and communication practices. The results provide unique evidence of the attention dynamics shaped by the research community today, including the dramatic growth and fast obsolescence of Computer Science eprints, which has not been captured in previous studies relying on the citation data of journal papers. Subsequently, we develop a quantitatively and temporally normalized citation index with an approximately normal distribution, which is useful for comparing citational attention across disciplines and time periods. Further, we derive a stochastic model consistent with the observed quantitative and temporal characteristics of citation growth and obsolescence. The findings and the developed framework open a new avenue for understanding the nature of citation dynamics.

1. INTRODUCTION

Scientists stand on the shoulders of giants by citing their predecessors' works in their own works[1]. Simultaneously, the scientific community often assumes, either explicitly or implicitly, that the more citations, the higher the scientific value. This assumption has long made the citation count and its various derivatives influential in research evaluation and administrative decision-making, offering practical, albeit imperfect, proxy metrics of scientific quality or importance (Garfield, 1972; Hirsch, 2005; Waltman, 2016). While such use of the citation-based metrics is controversial and often criticized (Aksnes, Langfeldt, & Wouters, 2019; Garfield, 1979, 2006; Lehmann & Lautrup, 2006; Line, 1993; Nicolaisen, 2007; Radicchi, Weissman, & Bollen, 2017; Seglen, 1998), they are nevertheless relevant to quantify the

[1] "*If I have seen further, it is by standing on the shoulders of Giants,*" Isaac Newton's letter to Robert Hooke, 1675.

an open access journal

Check for updates

Citation: Okamura, K. (2022). Scientometric engineering: Exploring citation dynamics via arXiv eprints. *Quantitative Science Studies*, *3*(1), 122–146. https://doi.org/10.1162/qss_a_00174

DOI: https://doi.org/10.1162/qss_a_00174

Supporting Information: https://doi.org/10.1162/qss_a_00174

Received: 26 June 2021
Accepted: 19 September 2021

Corresponding Author: Keisuke Okamura
okamura@ifi.u-tokyo.ac.jp

Handling Editor: Ludo Waltman

The MIT Press

attention of other scientists working on related issues, embodying the scientific knowledge transfer and information flow in the academic sphere. The life of scientific literature in terms of citations varies, experiencing widely different temporal patterns (He, Lei, & Wang, 2018; Ke, Ferrara et al., 2015; Redner, 2005; van Raan, 2004; Wang, Song, & Barabási, 2013). Some papers remain continually cited and even become "immortal giants," while others show fluctuating patterns of citation collection, and, in fact, a significant portion of papers apparently end their lives with little or no citational impact. Nevertheless, once aggregated and averaged across papers in a given research field, the average citation history curves typically follow a "jump–decay" pattern (Barnett, Fink, & Debus, 1989; Glänzel & Schoepflin, 1995); it increases up to a few years after publication to reach a peak and subsequently decreases to some asymptotic value.

The citation pattern has long fascinated researchers and practitioners alike. They have introduced the concept of literature obsolescence (aging) (Gross & Gross, 1927; Line, 1993; Line & Sandison, 1974)—the process of becoming no longer cited or less used—to investigate how quickly a scientific publication moves in and out of the focus of researchers' attention. Here, obsolescence does not necessarily mean deterioration in its intrinsic value but is a natural consequence of scientific development or technological advancement over time. Scientific discovery presented in a paper and the value it creates are eventually integrated into subsequent papers, thereby contributing to the shoulders to stand on. Having said that, an interesting question is whether the obsolescence rate has changed over time. Some studies (Larivière, Archambault, & Gingras, 2008; Martín-Martín, Orduna-Malea et al., 2016; Verstak, Acharya et al., 2014) have suggested that researchers are increasingly relying on older publications, while others (Evans, 2008; Parolo, Pan et al., 2015) have suggested that the obsolescence of scientific knowledge has accelerated over the past decades. Although results and implications from these previous studies vary without consensus (see also Pan, Petersen et al., 2018; Sinatra, Deville et al., 2015; Zhang & Glänzel, 2017), a common feature was the use of the citation data on papers published in peer-reviewed journals. The bibliometric data used were commonly obtained from publishers or major bibliographic databases, such as Web of Science[2] and Scopus[3].

Nowadays, the study of literature growth and obsolescence has embarked on an entirely new phase. This change is due to the new digital technologies, resources, and online interfaces that have enormously increased the speed and volume of scientific knowledge production and consumption. With the ongoing rapid and large-scale digitization of information, researchers increasingly rely on digital research documents, or eprints, available on a free and open-access platform. These eprints mainly include preprints, not yet peer-reviewed and accepted in scientific journals or other traditional means. Still, they are citable as they are publicly disclosed and accessible on the internet via the preprint servers, which have become an integral part of today's scholarly publishing system. arXiv[4], the most prominent and oldest eprint archive, launched in August 1991, has served as an indispensable research platform (document submission and retrieval system) for physics, mathematics, and computer sciences (arXiv.org, 2021; Ginsparg, 2016). For the past 30 years, it has functioned as a primary source of current and ongoing research, accelerating recognition and dissemination of research findings and facilitating rapid scholarly communication. The number of eprints

[2] Web of Science, published by Clarivate analytics or Thomson Reuters: https://clarivate.com/webofsciencegroup/solutions/web-of-science/.

[3] Scopus, published by Elsevier: https://www.scopus.com.

[4] arXiv.org [eprint archive]: https://arxiv.org/.

posted on arXiv has grown dramatically, from 30,601 (2000) to 155,866 (2019), increasing more than five-fold during the past 20 years (Supplementary Figure S1a), and the number of citations received by the arXiv eprints has also snowballed with time (Supplementary Figure S1b). Previous studies have investigated how arXiv has impacted traditional publication practices and the researchers' citation activities (Aman, 2013; Feldman, Lo, & Ammar, 2018; Gentil-Beccot, Mele, & Brooks, 2010; Larivière, Sugimoto et al., 2014; Moed, 2007; Okamura, Yoda et al., 2020; Wang, Chen, & Glänzel, 2020). For instance, evidence has been provided that arXiv has an advantage in accelerating citation (Aman, 2013; Moed, 2007; Okamura, Yoda et al., 2020) (see also Supplementary Figure S2). Recently, the preprint mode of scholarly communication has seen an exponential rise during the COVID-19 global pandemic, driven by the vital need for early and rapid dissemination of COVID-19-related research results. Two growing preprint servers are bioRxiv[5] (launched in November 2013) and medRxiv[6] (launched in June 2019), which have offered indispensable platforms for researchers in biomedical disciplines (Abdill & Blekhman, 2019; Fraser, Momeni et al., 2020; Fraser, Brierley et al., 2021; Fu & Hughey, 2019; Kirkham, Penfold et al., 2020; Sevryugina & Dicks, 2021). Scholarly communications in today's digitalized academic sphere have thus been rapidly shifting from the conventional journal-centered publishing styles toward a new mode of communication via preprint servers and social media (e.g., blogging, Facebook, and Twitter), per Open Science practices (Berg, Bhalla et al., 2016; Fraser et al., 2021; Gentil-Beccot et al., 2010; Ginsparg, 2016; Shuai, Pepe, & Bollen, 2012; Thelwall, Haustein et al., 2013; Wang et al., 2020).

Facing this publishing paradigm shift, what constitutes the growth and obsolescence of scientific knowledge has also been changing and diversifying. In many research areas and subfields, nowadays, it is quite common that eprints posted on preprint servers or repositories are not yet or never to be published in journals indexed in the commercial bibliometric databases (Abdill & Blekhman, 2019; Larivière et al., 2014; Okamura, Yoda et al., 2020; Sevryugina & Dicks, 2021). Even if the preprints are to be published eventually in a database-indexed journal, the citations received during the preprint duration (in their preprint forms) may not be fully covered in major bibliometric databases, leading to bias in the citation analysis. In addition, there exists a field-dependent bias resulting from the database coverage (Waltman, 2016). For instance, computer science has a unique conference-centric publishing culture, in which researchers have a tendency to value top-tier conferences as a publication venue rather than high-impact journals (Feldman et al., 2018; Kim, 2019). Also, the main research outputs in humanities and social sciences studies have been books rather than journal papers. For these disciplines, the citational impact of a scientific work tends to be underrepresented in many commercial bibliometric databases. Consequently, research methods and measurement based on the conventional publishing systems may often not be desirable to grasp the whole volume and dynamics of scientific attention, or citations, in this new era of Open Science.

With the aim of obtaining a better understanding of the real dynamics underlying the citation network beyond the traditional journal-centered scholarly communication, we investigate the citation data of eprints posted on arXiv. The data contain the information on various types of eprints, regardless of which journal (or none) eventually published (or will publish) them. A disadvantage of using the arXiv data compared to the commercial databases would be the limited variety in the subject categories; for instance, arXiv's research disciplines do not include biomedical disciplines. Still, the arXiv data cover various subject categories focusing on physics, mathematics, and computer sciences, and the long-term citation data (since 1991)

[5] bioRxiv [Preprint server]: https://www.biorxiv.org/.
[6] medRxiv [Preprint server]: https://www.medrxiv.org/.

provide a unique test bed for modeling and visualizing the distinct disciplinary patterns of scientific attention. Making full use of the advantages, we investigate two dimensions of citation dynamics: the "quantitative" dimension and the "temporal" dimension. The quantitative dimension regards the degree distribution in the citation network (Price, 1965, 1976). We consider both the power law model (Peterson, Pressé, & Dill, 2010; Price, 1976; Redner, 1998) and the lognormal model (Radicchi, Fortunato, & Castellano, 2008; Redner, 2005; Sheridan & Onodera, 2018), and discuss the extent to which each model explains the empirical data. We conclude that, by and large, the arXiv citation data obey the lognormal law. The other, temporal dimension concerns the time dependence of citation accumulation at the discipline-average level (Barnett et al., 1989; Glänzel & Schoepflin, 1995; Gross & Gross, 1927). The entire span of the growth and obsolescence pattern is modeled by a simple nonlinear function of time, which fits the empirical data remarkably well for all disciplines.

The revealed characteristics of the citation dynamics open up some interesting research directions; we discuss two directions in this paper. The first direction concerns the "fair" evaluation of scientific attention cast on individual papers. Although citation-based metrics have increasingly been applied in research evaluation and science policymaking, various sources of bias have limited their validity and utility (Aksnes et al., 2019; Lehmann & Lautrup, 2006; Seglen, 1998; Waltman, 2016). A particularly significant issue is the disciplinary difference, including the size of the research community, the typical time to publication, the average level of per-paper citations, the growth and obsolescence rates, and the degree to which the preprint (eprint) mode of communication is adopted[7]. The knowledge of citation dynamics revealed in this paper can mitigate some of these biases in both the quantitative and temporal dimensions, through a new index of citational attention called the γ-index, as will be demonstrated. The second direction is more theory oriented, which is an effort to reveal the underlying mathematical model that consistently explains the observed evolution pattern of citations, both quantity-wise and time-wise. Unlike the commonly accepted model of preferential attachment (Barabási & Albert, 1999; Barabási, Jeong et al., 2002; Newman, 2001), which generates networks with a power law, our stochastic model naturally reproduces the observed lognormal law in the quantitative dimension while also reproducing the observed evolution pattern ("jump–decay" plus "constant attention") of the per-paper average citations in the temporal dimension. We hope this paper serves to provoke further discussions and investigations of the related issues, thereby shedding new light on the nature of collective attention dynamics on scientific knowledge.

2. THE QUANTITATIVE DIMENSION OF CITATION DYNAMICS

We first describe the bibliometric data used throughout this study; see Appendix A in the Supplementary Materials for the data analysis and visualization platforms. Subsequently, we analyze the degree distribution in the arXiv citation network, providing evidence for the lognormal law in the quantitative dimension of the citation dynamics.

2.1. The arXiv Data

Our bibliometric analyses were based on a data set of 1,589,006 eprints posted on arXiv from its launch in 1991 until the end of 2019—hereafter referred to as "the arXiv eprints"—plus the

[7] Other biases arising from the motives or limitations on the citer's side include reference copying, self-citations, negative citations, politically motivated flattery, cronyism, limited space for references, homographs, language, nationality, and bibliometric database coverage, to name but a few. See, for example, Seglen (1998), Waltman (2016), and Aksnes et al. (2019) for detailed discussions.

data on their citations. This raw data set is the same as that used in Okamura, Yoda et al., (2020), which can be found in the Zenodo repository (Okamura & Koshiba, 2021). Here, eprints include preprints, conference proceedings, book chapters, data sets and commentary (i.e., all electronic material that has been posted on arXiv). The content and metadata of the arXiv eprints were retrieved from the arXiv API[8] as of January 21, 2020, where the metadata included data about the eprint's title, author, abstract, subject category, and arXiv ID, which is arXiv's original eprint identifier. In addition, the associated citation data were derived from the Semantic Scholar API[9] from January 24, 2020 to February 7, 2020, containing the citation information in and out of the arXiv eprints and their published versions (if applicable). Here, whether an eprint has been published in a journal or other means is assumed to be inferrable, albeit indirectly, from the status of the digital object identifier (DOI) assignment. It is also assumed that if an arXiv eprint received c_{pre} and c_{pub} citations until the data retrieval date (February 7, 2020) before and after it is assigned a DOI, respectively, then the citation count of this eprint is recorded in the Semantic Scholar data set as $c_{pre} + c_{pub}$. Both the arXiv API and the Semantic Scholar data sets contained the arXiv ID as metadata, which served as a key variable to merge the two data sets.

Our classification of research disciplines was based on that described in the arXiv.org (2021) website. There, the 153 arXiv subject categories are aggregated into several disciplines, of which we restrict our attention to the following six disciplines: Astrophysics (*astro-ph*), Computer Science (CS) (*comp-sci*), Condensed Matter Physics (*cond-mat*), High Energy Physics (HEP) (*hep*), Mathematics (Math) (*math*) and Other Physics (*oth-phys*), which collectively accounted for 98% of all the eprints[10]. Those eprints that are tagged to multiple arXiv disciplines (Supplementary Figures S3a, b) were counted independently for each discipline. Due to this overlapping feature, the final data set contained a cumulative total of 2,011,216 eprints. See Table 1 for the numbers of eprints by discipline (see Supplementary Table S1 for a detailed description). A notable difference was observed between the four physics-based disciplines and the nonphysics disciplines (CS and Math) in the distribution of the time elapsed without being assigned a DOI (Supplementary Figure S4), in line with the findings of previous studies (Aman, 2013; Larivière et al., 2014). Specifically, while the majority of physics-based eprints acquire a DOI within approximately 2 years and exhibit a steep slope down toward the third year, the majority of nonphysics-based eprints are never assigned a DOI (CS: 77%, Math: 69%) to the data retrieval date, exhibiting a heavy-tailed histogram. This is highly suggestive of the empirical fact that journal publication is not necessarily the venue of choice for researchers of CS and Math, and that they heavily rely on arXiv for tracking new research findings (Kim, 2019; Wang et al., 2020).

2.2. Degree Distribution of the Citation Network

We begin with the investigation of the degree distribution in the arXiv citation network. This field of study has a long history, beginning with the pioneering work by Price (1965), who observed that the citation distribution follows a power law. Following that, previous studies commonly considered either a (variation of) a power law (Eom & Fortunato, 2011; Hajra & Sen, 2006; Peterson et al., 2010; Price, 1976; Redner, 1998; Thelwall, 2016) or a lognormal law (Eom & Fortunato, 2011; Radicchi et al., 2008; Redner, 2005; Thelwall, 2016) to

[8] arXiv API: https://arxiv.org/help/api/. Accessed January 2020.
[9] Semantic Scholar API: https://api.semanticscholar.org/. Accessed January 2020.
[10] The remaining eprints included those categorized in Electrical Engineering and Systems Science, Statistics, Quantitative Biology, and Quantitative Finance.

Table 1. Classification of the arXiv disciplines. The classification scheme is based on that used on the arXiv.org (2021) website (a detailed description is presented in Supplementary Table S1).

Discipline	No. of eprints	arXiv subject classification
astro-ph	257,864	Astrophysics
comp-sci	386,507	Computer Science
cond-mat	263,506	Condensed Matter Physics
hep	286,840	High Energy Physics (Theory, Phenomenology, Lattice, Experiment)
math	428,621	Mathematics (including: Mathematical Physics)
oth-phys	387,878	Other Physics (including: Nuclear Theory and Experiment, General Relativity and Quantum Cosmology, Quantum Physics, Nonlinear Sciences)

characterize the citation distribution of journal papers. Here, we investigate the citation data on the arXiv eprints and verify the fitness of each model. Let $c_k \in \mathbb{Z}_0^+$ be the number of citations that an eprint labeled by k (hereafter referred to as *eprint* K) accumulates. As Figure 1 indicates, the distribution of the raw citations is extremely skewed toward very highly cited eprints (note that the plots are shown in the log-log scale; see also Supplementary Table S2). Given that, let us define

$$y_k \coloneqq \ln(c_k + 1) \tag{1}$$

with the "(+1)-shifted" citation variable (Sheridan & Onodera, 2018; Thelwall, 2016). By this definition, y is a monotonically increasing function of the raw citation variable, c, with the zero-citation ($c = 0$) case being mapped to $y = 0$. Also, let q_k represent the quantile rank of c_k in the citation distribution (i.e., $q_k = P(c \leq c_k)$). For the power law model with exponent $a > 1$,

$$P_{\mathrm{PL}}(c) \sim (a-1)(c+1)^{-a}, \quad \text{i.e.,} \quad y_{\mathrm{PL}} \sim \frac{\ln(1-q)}{1-a}, \tag{2}$$

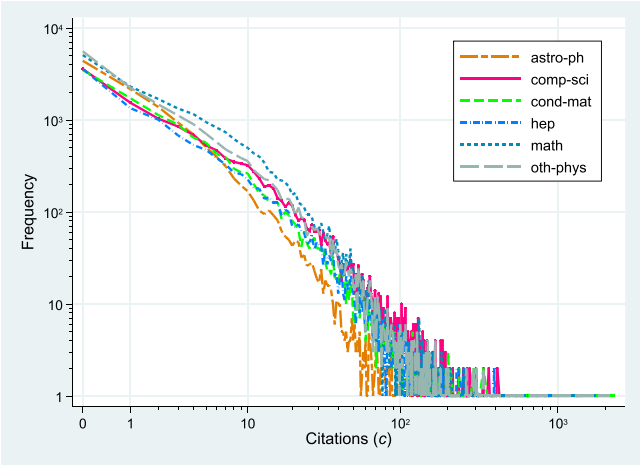

Figure 1. Distribution of cumulative citations of the arXiv eprints. Frequency distributions of citations to the eprints posted on arXiv in 2010 are shown in the log-log scale by discipline.

while for the lognormal model with mean b and variance m^2,

$$P_{\mathrm{LN}}(c) \sim \frac{1}{(c+1)m\sqrt{2\pi}} \exp\left[-\frac{(\ln(c+1)-b)^2}{2m^2}\right], \quad \text{i.e.,} \quad y_{\mathrm{LN}} \sim b + m\Phi^{-1}(q), \tag{3}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function defined by $\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-t^2/2}\, dt$.

To investigate which (or both, or neither) of the models (y_{PL} and y_{LN}) explains the empirical distribution of the arXiv citation data better, we created the quantile plot of y_k against each model, using the subdata sets of eprints classified according to the year of first submission (1991–2019). The fitness of each model was checked by visual inspection of the linearity between y_k and $\ln(1 - q_k)$ (power law; Supplementary Figure S5), and between y_k and $\Phi^{-1}(q_k)$ (lognormal; Supplementary Figure S6). First, let us look into the results for the power law model; see Figure 2a for the quantile plot based on the year-2010 data set. An approximately linear relationship can be seen in the region $-\ln(1 - q) \gtrsim 4$, or $q \gtrsim 0.98$ in terms of the quantile rank. However, this model does not fit well with the vast majority of the observations (i.e., below approximately the 98th percentile eprints). To cure this situation, we also considered a shifted power law model (Eom & Fortunato, 2011; Thelwall, 2016), $P_{\mathrm{sPL}}(c) \sim ((a-1)/\theta)(c/\theta + 1)^{-a}$, by introducing a shift parameter $\theta > 0$ (Supplementary Figure S7). Indeed, the approximately linear region could be expanded by fine-tuning the level of θ (see, e.g., the plot with $\theta = 10$). However, without any justification for its dynamical origin, the shift parameter (θ) represented nothing more than an arbitrary threshold set by hand. Also, the value of θ was quite sensitive to the slope of the linear region. Consequently, the estimated value of the exponent (a) was crucially dependent on the choice of the arbitrary parameter.

By contrast, the lognormal model was shown to fit the empirical data remarkably well; see Figure 2b for the quantile plot based on the year-2010 data set. An approximately linear relationship between y_k and $\Phi^{-1}(q_k)$ is observed for the entire region of $c_k > 0$, without any arbitrary thresholds as in the shifted power law model case. Applying an ordinary least squares regression model to the nonzero-citation data, the two lognormal parameters (b, m) were estimated with a very high coefficient of determination (the adjusted R-squared > 0.99; Supplementary Table S3), with which the degree distribution was estimated as $y \sim \mathcal{N}(\hat{b}, \hat{m}^2)$. This feature was also true for each data set year and each discipline (Supplementary Figure S6). These results lead us to conclude that the arXiv citation distribution is better explained by the lognormal model than the power law model[11]. Note that for both the power law and the lognormal models, the empirical data deviate from the law in the region $1 - q \lesssim \mathcal{O}(10^{-5})$ (i.e., exceptionally highly cited eprints).

3. THE TEMPORAL DIMENSION OF CITATION DYNAMICS

Next, we move on to the temporal dimension of the citation dynamics. As the following argument applies equally to journal papers and eprints, we use the term *paper* instead of *eprint* for the moment.

[11] Other skewed distributions such as the gamma and Weibull distributions, including the exponential distribution as their special case, were also considered. However, the lognormal distribution outperformed the others in describing the empirical arXiv citation data.

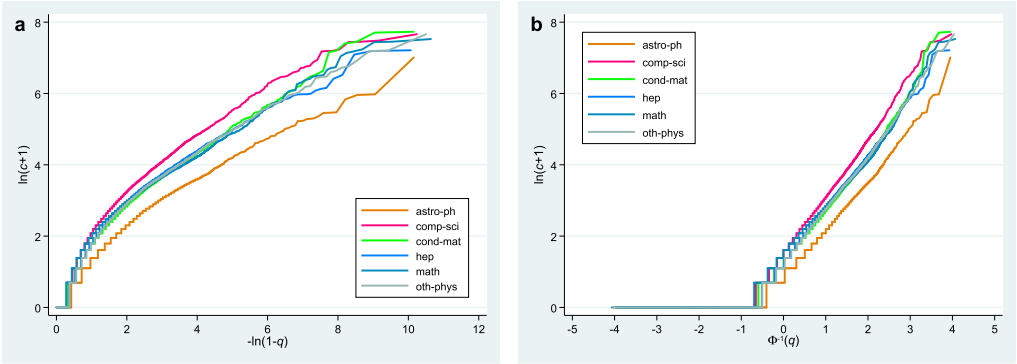

Figure 2. Degree distribution of the arXiv citation network: "Power Law vs. Lognormal Law." **a**: Plot of $y_k = \ln(c_k + 1)$ against $-\ln(1 - q_k)$, where c_k is the number of citations eprint k accumulates, and q_k is the quantile rank of c_k in the citation distribution of each discipline. **b**: Plot of y_k against $\Phi^{-1}(q_k)$, where $\Phi(\cdot)$ represents the standard normal cumulative density function. The plots are based on the citation data of eprints posted on arXiv in 2010.

3.1. Discipline-Average Citation History Curves

There have been two approaches known in the literature to investigate the scientific literature growth and obsolescence. The first approach is called the *retrospective* (or *synchronous, backwards-looking, citations from*) approach (Barnett et al., 1989; Burrell, 2002; Glänzel, 2004; Larivière et al., 2008; Nakamoto, 1988; Pan et al., 2018; Price, 1965; Redner, 2005; Verstak et al., 2014; Yin & Wang, 2017; Zhang & Glänzel, 2017), which looks at the distribution of the age of papers cited by papers published in a given year. The second approach is called the *prospective* (or *diachronous, forward-looking, citations to*) approach (Bouabid, 2011; Bouabid & Larivière, 2013; Burrell, 2002; Glänzel, 2004; Nakamoto, 1988; Pan et al., 2018; Parolo et al., 2015; Redner, 2005; Wang et al., 2013; Yin & Wang, 2017), which looks at the distribution of citations acquired over time by papers published in a given year. Our methodology was based on the latter, prospective approach, investigating the citation history of papers until the data retrieval date. Also, our analyses here are conducted at the discipline-average level, in which the citation dynamics is driven by collective scientific attention. The discipline-average trajectory of citation accumulation—*average citation history curve*—is indicative of the typical growth and obsolescence patterns of papers associated with the research discipline.

We conceptualize and model the per-paper average citation history curve as the superposition of the three components:

$$\text{average citation history} \approx (i)\text{ "jump–decay"} + (ii)\text{ "constant attention"} + (iii)\text{ "anomalies."} \quad (4)$$

The first component, (*i*), represents the pattern in which the average citation count increases with time to its peak value and then decreases as additional time passes, shaping a positively skewed curve. The second component, (*ii*), is assumed to derive from those papers that continue to acquire citations even after a paper's typical lifespan (Bouabid, 2011; Bouabid & Larivière, 2013; Redner, 2005). The cumulative advantage (Barabási & Albert, 1999; Barabási et al., 2002; Newman, 2001; Price, 1976; Redner, 2005; Sheridan & Onodera, 2018) (i.e., a tendency that highly cited papers are more likely to accumulate additional citations than papers with fewer citations) would contribute to this component. Here, although the classic preferential attachment model (Barabási & Albert, 1999; Barabási et al., 2002; Newman, 2001) is not necessarily assumed in its original form, a variation of the Matthew

effect (*the rich get richer*) (Merton, 1968) at the paper level is expected to be in operation. The third component, (*iii*), arises from unexpected or unpredictable citation events, accounting for deviations from the "regular" part of the curve, (*i*) + (*ii*). This anomalous component includes cases of multiple humps and the Sleeping Beauties in science (He et al., 2018; Ke et al., 2015; Redner, 2005; van Raan, 2004), whose citation history exhibits a long unrecognized period followed by sudden and intense attention.

As the number of papers increases, the contributions from this anomalous component, (*iii*), tend to be suppressed, and the average citation history curve will become approximately the regular part, (*i*) + (*ii*). However, cases of very highly cited papers, including some giant Sleeping Beauties, could still be impactful enough to deform the average citation history curve from the regular pattern. In fact, previous studies have found that highly cited publications tend to follow varied citation history curves (Redner, 2005; Wang et al., 2013). To mitigate such anomalous effects on our analysis of the average citation history curve, we set the threshold percentile of the citation distribution as the 99th percentile, and focus on the eprints below this threshold (hereafter referred to as *the below-99th percentile eprints* or *the below-99th citation data*)[12]. The rest, the top 1% highest cited eprints, were removed from analyses, as the average citation history curve of such a highest cited cluster of eprints no longer follows the regular time-dependence of citation accumulation (Supplementary Figures S8a, b).

3.2. Regression Model Fitting

Now we are in a position to construct a regression model for the time distribution of citations. Let u_i denote the average citations of eprints of i years old ($i \in \mathbb{Z}_0^+$). It is defined as the total citation counts acquired by eprints posted on arXiv i years ago, divided by the total number of eprints posted on arXiv i years ago. Note that, in general, the mean value is not regarded as a good measure of central tendency for a highly skewed distribution, in which case the use of the median is usually preferred. However, regarding the current arXiv data, the median yielded the values of one or two for all disciplines due to the extremely skewed distribution of citations (Supplementary Table S2), which were not useful for comparative analysis. It then turned out that the mean value served as a useful indicator to investigate the time profile of the citations, allowing for a comparative analytical approach. Moreover, the use of the mean value also dovetailed with the stochastic modeling of citation dynamics discussed later (Section 4.2). Let also t_i denote the age of eprints posted on arXiv i years ago (i.e., $t_i := i$). Then, our regression equation can be stated as

$$u_i = Af(t_i + 1; \mu, \sigma) + Bg(t_i; \lambda) + \epsilon_i, \tag{5}$$

where $A > 0$ and $B > 0$ are the overall scaling factors for parametric functions f and g (see below for the explicit expressions), respectively, and ϵ_i is the error term. This model function comprises two parts. The first component,

$$f(t; \mu, \sigma) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left[-\frac{(\ln t - \mu)^2}{2\sigma^2}\right], \tag{6}$$

captures the jump–decay pattern, (*i*), of Eq. 4. It is represented by a lognormal probability distribution function so that $\ln t \sim \mathcal{N}(\mu, \sigma^2)$. Here, we note that previous studies have considered a

[12] More generally, we refer to the eprints at and below the pth percentile of the citation distribution as *the below-pth percentile eprints or the below-pth citation data*.

variety of functional forms to model the obsolescence (aging) process in citations, including power law (Hajra & Sen, 2005; Parolo et al., 2015), exponential (Hajra & Sen, 2006; Parolo et al., 2015), polynomial (Bouabid, 2011; Bouabid & Larivière, 2013) and lognormal (He et al., 2018; Wang et al., 2013; Yin & Wang, 2017), of which some are practical or heuristic, and some are theory based. Of all these alternative specifications of the model function, we employed the lognormal function not only for its intuitive clarity and neatness but also for the empirical and theoretical soundness (see Appendix B in the Supplementary Materials for a theoretical validation of the lognormal time distribution). Some basic statistical parameters characterizing this distribution are obtained from the two parameters, μ and σ, as: Mean = $e^{\mu+\sigma^2/2}$, Median = e^{μ}, Mode = $e^{\mu-\sigma^2}$, and Variance = $(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$. Note that the argument of f in Eq. 5 is set as $t_i + 1$, rather than t_i, so that the contribution of the lognormal component to the total citation outcome at $i = 0$ becomes $Af(t_i + 1; \mu, \sigma)|_{i=0} = A(\sqrt{2\pi}\sigma)^{-1}\ e^{-\mu^2/(2\sigma^2)}$. The second component in Eq. 5,

$$g(t;\lambda) = \tanh(\lambda t), \tag{7}$$

represents a monotonic increasing, sigmoid function with parameter $\lambda > 0$ so that $g(t = 0; \lambda) = 0$ and $g(t; \lambda)|_{t\gg 1} = 1$. This component accounts for the constant attention component, (*ii*), of Eq. 4. Here, we employed the hyperbolic tangent as the model function for its simplicity while acknowledging other sigmoid-type alternatives. The rationale of using a "soft" sigmoid function rather than a "hard" Heaviside step function—the large-λ limit of Eq. 7—is that the rate of rising can vary among disciplines (Wang, 2013). Specifically, for some disciplines, the constant attention comes into effect with an almost instant rise, while for others, it can take a few years before reaching a plateau. Note that the argument of g in Eq. 5 is set as t_i rather than $t_i + 1$, because the long-lasting citation effect can only come into effect after an eprint is posted on arXiv, so that $g(t_i; \lambda)|_{i=0} = 0$ for arbitrary λ. The three parameters (μ, σ, and λ) and the two overall factors (A and B) were estimated by applying the nonlinear regression model, Eq. 5, to the yearly citation data for each discipline[13]. Thus, in contrast to most previous studies that treat the growth and obsolescence phases separately, often focusing on the latter, we quantify the discipline-average characteristics of the entire citation history.

3.3. Revealed Quantitative and Temporal Characteristics of Citation Evolution

Figure 3 shows the time distribution of the empirical data, $\{(t_i, u_i)|\, i = 0, \ldots, 20\}$, overlaid by the fitted regression curves based on Eq. 5 (i.e., $\hat{u}(t) = \hat{A}f(t; \hat{\mu}, \hat{\sigma}) + \hat{B}g(t; \hat{\lambda})$, $0 \le t \le 20$), for each arXiv discipline and for the below-pth percentile eprints ($p = 99, 95, 90, 75, 50$). Here, the estimates of parameters are indicated by a "hat" symbol. The coefficient of determination (the adjusted R-squared) was higher than 0.99 for all disciplines. Note that, considering the extremely skewed distribution of citations (Figure 1), the vertical axis may be better thought of as shown by an arbitrary scale, which nevertheless is useful for a cross-discipline comparative analysis. A wide variety of differences can be seen among the fitted curves across both disciplines and percentile sections. For each discipline panel, the vertical gaps between the regression curves reflect the variance of the quantitative citation distribution (cf. Supplementary Figure S9). Regarding the percentile section, focusing on the below-99th citation data, Table 2

[13] The nonlinear regression analyses were conducted using STATA/IC software (version 13; StataCorp LP, Texas, USA) via the `-nl-` command, which fits an arbitrary nonlinear function by least squares. https://www.stata.com/manuals13/rnl.pdf.

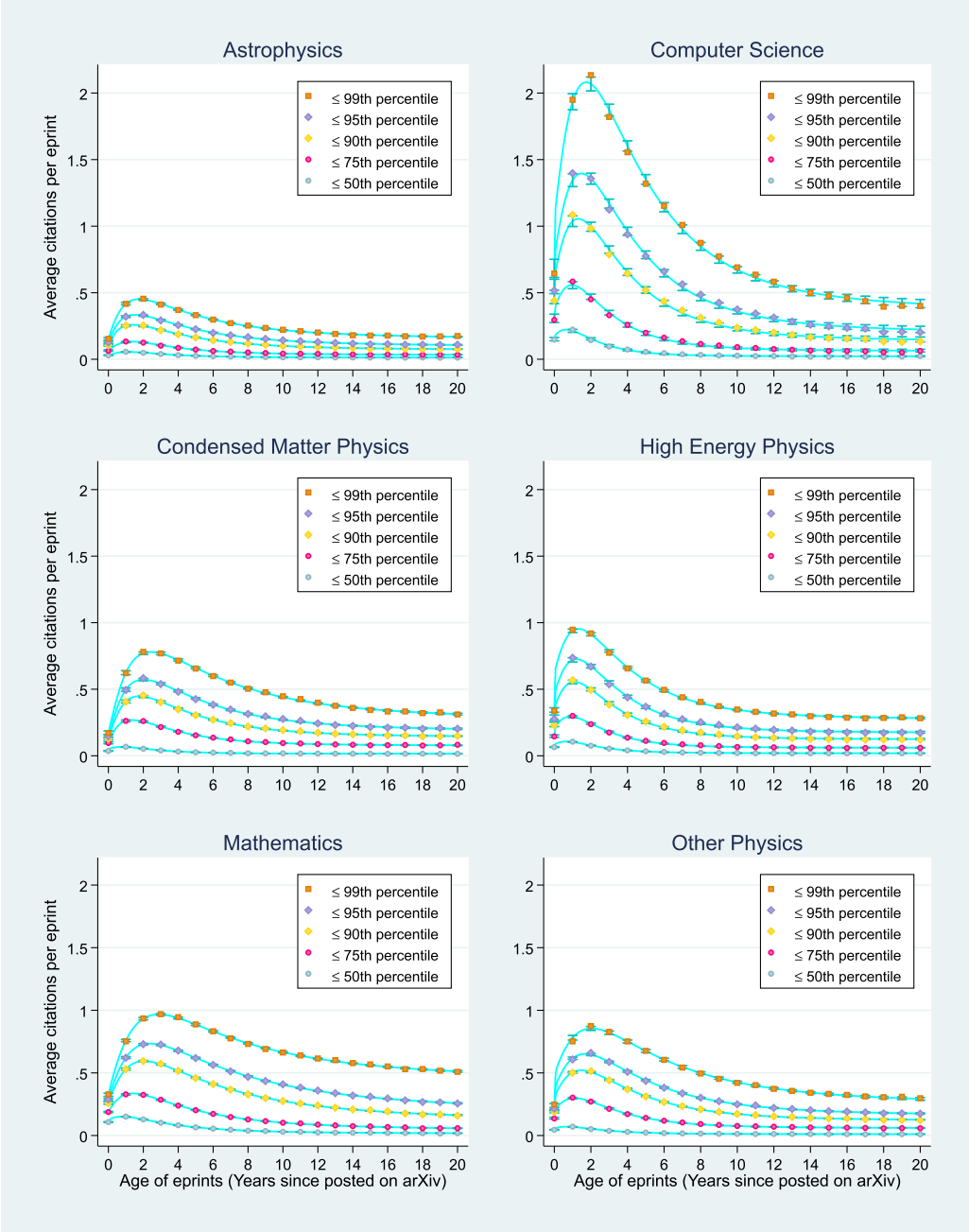

Figure 3. The citation distribution and the fitted regression curves by discipline. Symbols indicate the observations in the arXiv data at each percentile of the citation distribution. The curves represent the model function of Eq. 5 with the estimated parameters given in Table 2 (for the below-99th citation data) and Supplementary Table S4 (for the citation data with the lower percentile thresholds). The 99% confidence intervals are indicated by capped vertical lines.

summarizes the estimated regression coefficients with the derivative metrics, and Figure 4 shows the predicted discipline-average citation history curves (see Supplementary Table S4 and Supplementary Figure S10 for the results regarding the lower percentile data). As can be seen, CS exhibits the highest peak in a short time interval, indicating the highest growth rate of citation counts; its peak value ($\hat{u}_p = 2.08$) is more than double that of any other discipline. Also, Math exhibits the heaviest-tailed curve profile with the highest constant attention

Table 2. Estimated regression coefficients with the derived metrics. Results are shown for the below-99th citation data. The adjusted *R*-squared was higher than 0.99 for all disciplines. In addition to the estimated regression coefficients (left half), the peak value ($\hat{u}_p$), the typical time interval of the growth phase (δ_1), and the obsolescence phase (δ_2), the internal obsolescence rate ($\mathcal{S} = \delta_1/\delta_2$) and the retention rate ($\mathcal{R} = \hat{B}/\hat{u}_p$) (right half) are also shown by discipline. Values of $\hat{\lambda}$ larger than 10 are displayed as "$\gg 1$." (Results regarding the lower percentile data are presented in Supplementary Table S4).

Category	$\hat{A}$	$\hat{\mu}$	$\hat{\sigma}$	$\hat{B}$	$\hat{\lambda}$	$\hat{u}_p$	δ_1	δ_2	$\mathcal{S}$	$\mathcal{R}$
astro-ph	2.19	1.61	0.817	0.158	1.21	0.450	2.56	2.43	1.05	0.351
comp-sci	1.14×10	1.56	0.741	0.379	$\gg 1$	2.08	2.74	2.00	1.37	0.182
cond-mat	4.60	1.83	0.802	0.279	0.916	0.779	3.26	2.95	1.11	0.359
hep	3.71	1.37	0.725	0.277	$\gg 1$	0.953	2.32	1.61	1.45	0.290
math	6.25	1.91	0.927	0.452	0.439	0.918	2.85	3.88	0.735	0.493
oth-phys	5.04	1.76	0.805	0.259	$\gg 1$	0.855	3.03	2.77	1.10	0.303

component ($\hat{B}$ = 0.452)[14]. These characteristics are in sharp contrast to the remaining four physics-based disciplines. Differences among the physics-based disciplines are also evident concerning the peak height, the mode year, and the degree of stretching in the time direction. For example, Astrophysics has the lowest peak value ($\hat{u}_p$ = 0.450) and the lowest constant attention ($\hat{B}$ = 0.158). Also, the jump–decay phase of HEP occurs in a much shorter time interval than the other physics-based disciplines. All these characteristics reflect different citation, publication, and communication practices intrinsic to each research discipline (Björk & Solomon, 2013; Schubert & Braun, 1996; Zitt, Ramanana-Rahary, & Bassecoulard, 2005).

Besides, the timing in which the constant attention component—the second term in Eq. 5—comes into effect also varies across disciplines, from an almost instant rise (e.g., CS and HEP) to a gradual rise (Math). To elaborate on this point, let us decompose each predicted citation history curve into its components (i.e., the jump–decay (lognormal) component and the constant attention (sigmoid) component). Let $\rho(T)$ be the ratio of the citation contribution from the lognormal part to the total citations during the period $[0, T]$. Introducing the collective shorthand notations, $\omega_1 = \{A, \mu, \sigma\}$, $\omega_2 = \{B, \lambda\}$ and $\Omega = \{\omega_1, \omega_2\}$, it is evaluated by

$$\rho\left(T; \hat{\Omega}\right) \coloneqq \frac{F(T; \hat{\omega}_1)}{H\left(T; \hat{\Omega}\right)} = 1 - \frac{G(T-1; \hat{\omega}_2)}{H\left(T; \hat{\Omega}\right)}, \tag{8}$$

where F, G, and H are the cumulative functions defined by

$$F(T; \omega_1) = A \int_0^T f(t; \mu, \sigma) dt = A\Phi\left(\frac{\ln T - \mu}{\sigma}\right),$$

$$G(T; \omega_2) = B \int_0^T g(t; \lambda) dt = \frac{B}{\lambda} \ln(\cosh(\lambda T)), \tag{9}$$

and $H(T; \Omega) = F(T; \omega_1) + G(T-1; \omega_2)$, respectively. The time behavior of $\rho(T)$ varies across disciplines (Supplementary Figures S11 and S12). The CS, HEP, and Other Physics curves exhibit a cusp soon after the posting on arXiv, reflecting the rapid saturation (i.e., $\hat{\lambda} \gg 1$).

[14] A caveat in interpreting the result for Math is that around 10–20% of eprints (submission year: 2010–2019) categorized in the Math discipline were also cross-listed in the CS discipline (cf. Supplementary Figure S3b), which has grown exponentially in the past decades. Therefore, the quantitative characteristics of Math discussed in this study may not be directly compared to the findings of other work based on different data sources (e.g., Larivière et al., 2014; Wang et al., 2020).

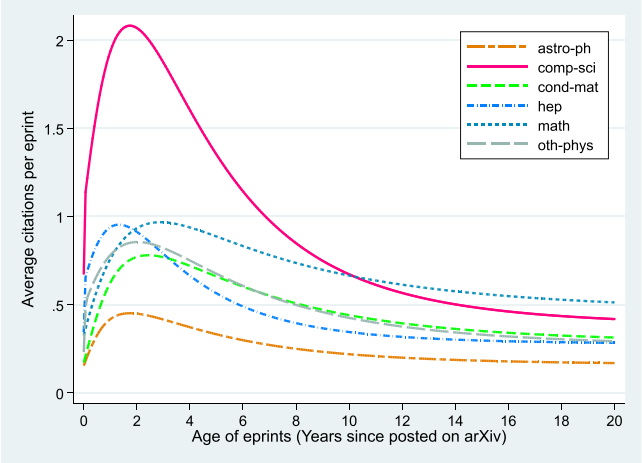

Figure 4. Discipline-average citation history curves for the arXiv disciplines. Fitted curves are based on the regression results for the below-99th percentile eprints. See Table 2 for the summary statistics.

As the age increases, ρ tends to be independent of the shape parameters ($\hat{\mu}$, $\hat{\sigma}$, $\hat{\lambda}$) and becomes dependent only on the overall factors ($\hat{A}$, $\hat{B}$). Specifically, $\rho \sim (1 + (\hat{B}/\hat{A})T)^{-1}$ as $T \gg 1$. Throughout the period (except for the first few years of the cusp period), ρ_{CS} remains at a high level ($\gtrsim$ 60%) for all percentile sections, reflecting the largest overall factor for the lognormal component ($\hat{A}$ = 11.4). By contrast, ρ_{Math} is the lowest for the below-99th citation data, but the highest for the below-50th citation data for $T \in [5, 10]$; the lower the percentile threshold, the higher the level of ρ_{Math}.

3.4. Speed of Obsolescence and Memory of Science

Having obtained the average citation history curve with quantitative characteristics, we now conduct a cross-discipline comparative analysis of the obsolescence rate. To characterize the obsolescence phenomenon, previous studies commonly used measures such as "half-life" (which often also refers to the median citation age) (Burton & Kebler, 1960; Larivière et al., 2008; Parolo et al., 2015; Zhang & Glänzel, 2017), average citation age (Larivière et al., 2008; Redner, 2005; Zhang & Glänzel, 2017), "life-expectancy" (or "life-time") (Bouabid, 2011; Bouabid & Larivière, 2013), and Price Index (Larivière et al., 2008; Zhang & Glänzel, 2017)[15]. An alternative way to conceptualize the obsolescence rate may be to use the magnitude of the down-slope (gradient) of the citation history curve after its peak, which involves both the quantitative and temporal dimensions. Whichever of the above indicators is used, it is subject to discipline-specific biases arising from the disciplinarily different publication practices, systems and processes (Björk & Solomon, 2013; Schubert & Braun, 1996; Zitt et al., 2005). For instance, regarding the quantitative bias, 10 citations received by a Life Science paper would not represent the same citational impact as that received by a Math paper (e.g., Schubert & Braun, 1996, Table 1; Okamura, 2019, Supplementary Table S2). This means that the drop in the raw number of yearly citations would not be a well-suited indicator of obsolescence. Also, regarding the temporal bias, what a year means to the academic life-cycle can be very different between disciplines. For instance, some disciplines have a high turnover rate of citation with a fast-paced research environment compared to other disciplines, as

[15] The "half-life" (median citation age) and the average citation age are defined in either a retrospective/synchronous or a prospective/diachronous manner, depending on the context.

already evidenced in our arXiv data (Figure 4 and Supplementary Figure S10). Indeed, an investigation of the predicted cumulative citation history curves revealed that, on average, more than half of the citations throughout the life-to-date of an HEP eprint has already been received within the first 3 years since posted on arXiv, whereas the corresponding figure for a Math eprint is below 30%. Besides, publishing delay also varies across disciplines (Aman, 2013; Björk & Solomon, 2013). Therefore, a comparative analysis of the obsolescence rates can be conducted only after first gauging the relevant bibliometric coordinates against a quantitatively and temporally normalized reference standard.

To achieve this goal, we consider the quantitative and temporal aspects of obsolescence separately and introduce two kinds of obsolescence metrics defined for each dimension. First, we investigate the time-wise obsolescence by focusing on the temporal characteristics of the lognormal (jump–decay) component of the average citation history curve. We introduce a discipline-specific time-scale (i.e., a typical unit time associated with each discipline). Just as the biological age—measured by a biological clock—is more relevant than the chronological age in some biological studies, here, "bibliodynamical" age—measured by a "bibliodynamical" clock—is more relevant to measure the obsolescence perceived by the researchers of each discipline. Let δ_1 and δ_2 denote the typical time interval of the growth phase and the obsolescence phase associated with a citation history curve. Among several possible ways to identify these characteristics in the lognormal distribution[16], we adopt the definitions that $\delta_1 :=$ Mode and $\delta_2 :=$ Median − Mode. These metrics can be directly obtained from the estimated regression parameters of the lognormal distribution, yielding $\delta_1 = e^{\hat{\mu}-\hat{\sigma}^2}$ and $\delta_2 = e^{\hat{\mu}}(1 - e^{-\hat{\sigma}^2})$. Subsequently, we define the time-adjusted rate of temporal obsolescence as the ratio of δ_1 to δ_2,

$$\mathcal{S} := \frac{\delta_1}{\delta_2} = \frac{1}{\exp(\hat{\sigma}^2) - 1}, \tag{10}$$

which we call the *internal obsolescence rate*. Literally, it represents how long it takes to reach the peak time (Mode) after being posted on arXiv, divided by how long it takes to reach the typical point in time lying in the skirts of the curve (Median) after the peak time. This quantity is dimensionless, depending only on the shape parameter, $\hat{\sigma}$, without dependence on the scale parameter, $\hat{\mu}$, of the lognormal distribution. Also, it behaves as $\mathcal{S}(\hat{\sigma}) \sim \hat{\sigma}^{-2}$ for $\hat{\sigma} \ll 1$, indicating a very fast (almost instant) obsolescence, and as $\mathcal{S}(\hat{\sigma}) \sim e^{-\hat{\sigma}^2}$ for $\hat{\sigma} \gg 1$, indicating a very slow (no) obsolescence.

It is intriguing to know how the obsolescence rate has changed over time (Evans, 2008; Larivière et al., 2008; Martín-Martín et al., 2016; Pan et al., 2018; Parolo et al., 2015; Verstak et al., 2014; Zhang & Glänzel, 2017). The findings from the previous studies have been inconclusive or inconsistent, as they use different sources of the citation data with different field categorizations and time windows. We revisited this issue by using the arXiv data with the proposed internal obsolescence rate of Eq. 10. We constructed a series of 10 consecutive subdata sets from the entire arXiv data set such that the first, the second, and the nth subdata set respectively contain the eprints posted on arXiv before and including the year 2019, 2018, and 2020 − n, with n running from 1 to 10. Subsequently, the regression model of Eq. 5 was applied to each subdata set. The estimated regression coefficients were used to obtain δ_1 and δ_2 for each data set year (2010–2019), with which the internal obsolescence rate

[16] Alternative ways to quantify δ_2, based on the lognormal characterization, include $\delta_2' =$ Mean − Mode and $\delta_2'' =$ Mean − Median, the resulting internal obsolescence rate simply being Eq. 10 with the term σ^2 replaced with $3\sigma^2/2$ and $\sigma^2/2$, respectively.

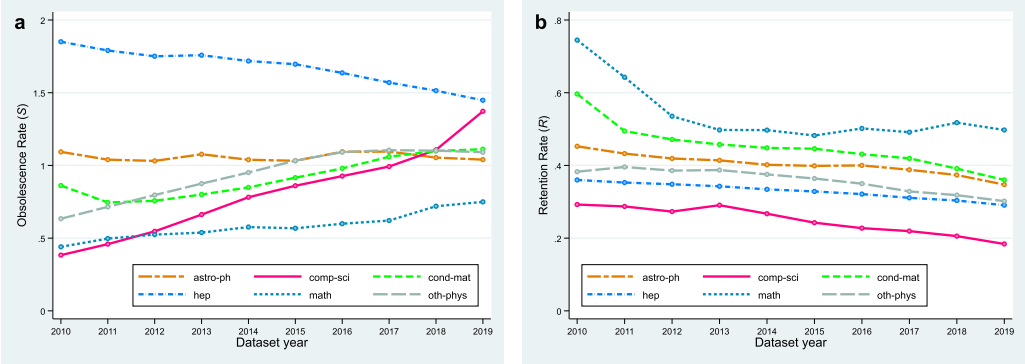

Figure 5. Trends in the internal obsolescence rate ($\mathcal{S}$) and the retention rate ($\mathcal{R}$). The analyses are based on the below-99th percentile eprints posted on arXiv during 2010–2019. **a**: Trends in the internal obsolescence rate defined by $\mathcal{S} = \delta_1/\delta_2$, where δ_1 and δ_2 respectively represent the typical time interval of the growth phase and the obsolescence phase identified in the average citation history curve. **b**: Trends in the retention rate defined by $\mathcal{R} = \hat{B}/\hat{u}_p$, where $\hat{B}$ and $\hat{u}_p$ respectively represent the asymptotic value and the peak value of the average citation history curve.

($\mathcal{S}$) is obtained by Eq. 10. Figure 5 shows the resultant trends in $\mathcal{S}$ by discipline. During the past decade, the most remarkable increase is observed for CS, for which $\mathcal{S}$ has increased from 0.383 (2010; the lowest of all) to 1.37 (2019; close to the highest). It implies that for CS researchers today, knowledge at the research front has increasingly become obsolete with the exponential growth in the CS eprints (Supplementary Figure S1). Condensed Matter Physics and Other Physics also follow an increasing trend with relatively similar levels of $\mathcal{S}$. Math shows a less steep increase from 0.440 (2010) to 0.749 (2019), remaining at the lowest level since 2012. Astrophysics stays relatively flat around slightly above $\mathcal{S} = 1$, suggesting that the speed of obsolescence does not vary much over time. By contrast, HEP exhibits a decreasing trend, for which $\mathcal{S}$ has steadily dropped from 1.85 (2010) to 1.45 (2019). Still, HEP kept at the highest level throughout the period, suggesting a highly competitive situation among researchers.

In contrast to the time-wise obsolescence discussed above, the quantity-wise obsolescence can be quantified as the degree to which the early peak level is retained after a while to reach the constant attention level. The corresponding metric is defined by

$$\mathcal{R} := \frac{\hat{B}}{\hat{u}_p}, \tag{11}$$

which we call the *retention rate*. By definition, it depends only on the (yearly) citation peak value and the asymptotic value, without dependence on the process or trajectory of the citation evolution. Using the same data sets as used for the internal obsolescence rates, the retention rates for the past years and the trend therein were also analyzed (Figure 5b). The ranking of the six arXiv disciplines has not changed over the period (2010–2019). CS, the lowest throughout, shows a steady decreasing trend, and the four physics-based disciplines also show a decreasing trend, indicating the gradual progress of the quantity-wise obsolescence. Math, the highest throughout, shows a significant drop during 2010–2013, but remains in a relatively flat trend around $\mathcal{R} = 0.5$ after 2013, suggesting that rather "old" Math eprints have continued to have a presence as a knowledge source. Finally, we note that the retention rate is reciprocally related to what could be called the *inflation rate*, defined by the ratio of the peak value, $\hat{u}_p$, to the baseline level, $\hat{B}$. Put differently, it measures the degree to which the early growth trend is inflationary compared to the stationary level reached after a while (see Supplementary Figure S13).

4. DISCUSSION

Based on the revealed knowledge of the quantitative and temporal dimensions of citation dynamics, this section extends our results in two directions. First, we develop a new citation-based index that can be used to compare the citational impact of papers published in different disciplines at different times. Second, we provide a mathematical (stochastic) model of citation dynamics that reproduces both the quantitative and temporal patterns observed in the previous sections. The first part will be of interest to practitioners of research evaluation and science policymakers who wish to improve the informed peer-review process, while the second part is mainly for interested researchers working on the intersection of scientometrics and mathematical physics/economics.

4.1. Quantitatively and Temporally Normalized Index of Citational Attention (γ-index)

Previous studies have developed a number of citation-based indices to quantify the research impact of scientific publications. These indices are often "field-normalized," reflecting the fact that the average citations per paper vary widely between fields (Schubert & Braun, 1996; Waltman, 2016; Zitt et al., 2005). However, the normalization scales used there have mostly been based on the bibliometric data on journal papers. Considering that scholarly communications, including citations, have already been or are being heavily dependent on eprint (preprint) systems in many research disciplines (see Section 1), the normalization scales are necessarily biased in both the quantitative and temporal dimensions without consideration of nonjournal-based citations[17]. In addition, an appropriate citation window varies between disciplines due to the different bibliodynamic clock discussed earlier. Towards a less biased evaluation and comparative analysis of the scientific attention, we propose a new, quantitatively and temporally normalized citation index based on the revealed characteristics of the average citation history curve.

Let $c_k(T)$ denote the number of cumulative citations of eprint k at some time point $t = T > 0$, posted on arXiv at $t = 0$. The above consideration regarding the quantitative and temporal biases leads us to introduce a proper normalization factor, which is precisely given by the cumulative density function introduced before (see the definition below Eq. 8), $H_{s(k)}(T) := H(T; \hat{\Omega}_{s(k)})$, evaluated for discipline s(k) to which eprint k belongs. The quantitatively and temporally normalized citation index for eprint k at time T is then defined by

$$\gamma_k(T) := \ln\left(\frac{c_k(T)}{H_{s(k)}(T)}\right), \tag{12}$$

which we call the γ-*index*. By this definition, $\gamma = 0$, $\gamma > 0$, and $\gamma < 0$ correspond to, respectively, a citation equal to, more than, and less than that expected by the discipline-average citation history curve. Here, we remind that $H_{s(k)}(T)$ has been estimated based on the mean values of the citation distribution, which is extremely skewed. In this sense, the γ-index may

[17] As an illustration, Supplementary Table S2 of Okamura (2019) reported the per-paper average citations (10-year average) for "Computer Science and Mathematics," "Physics & Space Sciences," and "Basic Life Sciences" to be 5.26, 11.9, and 14.7, respectively. A more detailed study (unpublished) by the same author also revealed that the corresponding figures for "CS," "Math," and "Physics" are 6.76, 4.37, and 11.2, respectively. These results were obtained from the Essential Science Indicators (ESI) database, published by Clarivate Analytics, which only covers journal-published papers. By contrast, the current analysis based on the arXiv data predicts that the per-eprint average citations for CS and Math as $H_{CS}|_{T=10} = 18.7$ and $H_{Math}|_{T=10} = 14.1$, respectively, which are far above those for the four physics-based disciplines, in sharp contrast to the journal-based findings. These differences apparently stem from the fact that research outputs in CS and Math are significantly underrepresented in the ESI database.

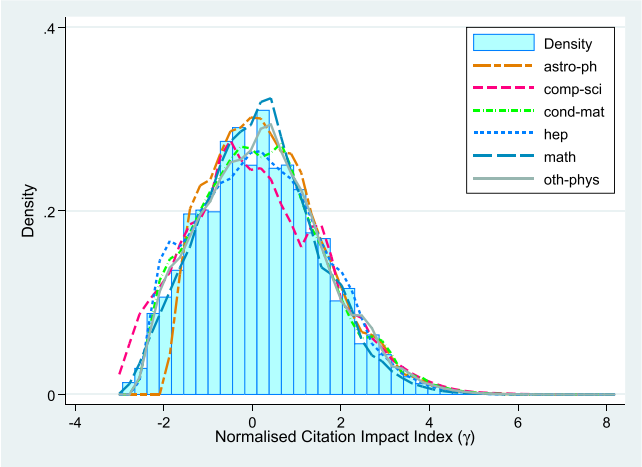

Figure 6. Distribution of the quantitatively and temporally normalized citation index (γ). Results are shown for the below-99th percentile eprints with nonzero citations. Shown is the frequency distribution of the γ-index, overlaid by the graphs of the kernel density estimation (half-width = 0.2) by discipline.

also be better thought of as quantified by an arbitrary scale. In what follows, we focus on the eprints with nonzero citations (hereafter referred to as *the nonzero citation data*; N = 1,206,997). Below, we restrict our analysis to those eprints posted on arXiv after and including the year 1999, which is the same set of eprints as used in the regression analysis before. Figure 6 shows the frequency distribution of the γ-index for the nonzero-citation data, overlaid by the graphs of the associated kernel density estimation by discipline (see also Supplementary Table S5). Overall, the γ-index tends to be distributed normally. However, the Gaussian approximation becomes invalid for the left tail of the distribution around $\gamma \lesssim -2$. This fact can also be verified through the lognormal quantile plot of γ (Supplementary Figure S14a). Also, the one-way analysis-of-variance (ANOVA) and the *post hoc* Bonferroni test for interdiscipline comparison showed that the mean values of the distribution of the γ-index were statistically different among disciplines ($p < 0.01$ for all pairs except $p = 0.169$ between Condensed Matter Physics and HEP).

To further refine the index, we also considered the associated standardized index defined by

$$\gamma_k^*(T) \coloneqq \Phi^{-1}(Q_k(T)), \tag{13}$$

where $Q_k(T)$ represents the quantile rank of γ_k at time T, restricted to the discipline s(k). This index, called the *γ^*-index*, has a distribution very close to the standard normal distribution (Supplementary Figure S14b, Supplementary Table S6). This time the one-way ANOVA and the *post hoc* Bonferroni test for interdiscipline comparison did not reject the null hypothesis that there was no difference among the disciplines ($p = 1.00$ for all pairs). This result supports the validity of the γ^*-index as the *standard* citation index, which can serve as a further less biased alternative to the existent citation-based impact measures. In practice, however, it is not always possible to calculate the quantile ranks due to limitations on the data availability. In that case, we can still use the γ-index of Eq. 12 as it can be obtained with only the knowledge of pre-estimated average citation history curves without the need to calculate the quantile ranks. Indeed, a high correlation between γ and γ^* was verified (Supplementary Figure S14a, Supplementary Table S7; the Pearson's correlation coefficient between the two indices was greater than 0.98 for all disciplines at the 0.1% significance level), suggesting that the γ-index is a good proxy for the γ^*-index.

Table 3. Ready reckoner for the normalized citation index (γ). Precalculated values are shown by discipline, based on the below-99th citation data. Negative γ-values are omitted and shown as a dash (—).

		$T = 2$	$T = 3$	$T = 4$	$T = 5$	$T = 6$	$T = 7$	$T = 8$	$T = 9$	$T = 10$
astro-ph	$c = 5$	2.61	1.82	1.39	1.12	0.92	0.77	0.66	0.56	0.48
	$c = 10$	3.31	2.51	2.08	1.81	1.61	1.47	1.35	1.26	1.17
	$c = 50$	4.92	4.12	3.69	3.42	3.22	3.08	2.96	2.87	2.78
	$c = 100$	5.61	4.82	4.39	4.11	3.92	3.77	3.65	3.56	3.48
comp-sci	$c = 5$	1.04	0.27	—	—	—	—	—	—	—
	$c = 10$	1.73	0.96	0.54	0.28	0.10	—	—	—	—
	$c = 50$	3.34	2.57	2.15	1.89	1.71	1.58	1.49	1.41	1.35
	$c = 100$	4.03	3.27	2.85	2.59	2.41	2.28	2.18	2.10	2.04
cond-mat	$c = 5$	2.35	1.43	0.93	0.61	0.38	0.21	0.08	—	—
	$c = 10$	3.04	2.13	1.62	1.30	1.08	0.91	0.77	0.67	0.57
	$c = 50$	4.65	3.74	3.23	2.91	2.69	2.52	2.38	2.27	2.18
	$c = 100$	5.35	4.43	3.93	3.61	3.38	3.21	3.08	2.97	2.88
hep	$c = 5$	1.68	0.98	0.61	0.37	0.21	0.09	—	—	—
	$c = 10$	2.38	1.68	1.30	1.06	0.90	0.78	0.68	0.61	0.54
	$c = 50$	3.99	3.29	2.91	2.67	2.51	2.39	2.29	2.21	2.15
	$c = 100$	4.68	3.98	3.60	3.37	3.20	3.08	2.99	2.91	2.84
math	$c = 5$	1.98	1.17	0.69	0.37	0.13	—	—	—	—
	$c = 10$	2.67	1.86	1.38	1.06	0.83	0.65	0.50	0.39	0.29
	$c = 50$	4.28	3.47	2.99	2.67	2.43	2.26	2.11	1.99	1.89
	$c = 100$	4.97	4.16	3.68	3.36	3.13	2.95	2.81	2.69	2.59
oth-phys	$c = 5$	1.94	1.17	0.74	0.45	0.25	0.10	—	—	—
	$c = 10$	2.63	1.86	1.43	1.15	0.94	0.79	0.67	0.57	0.49
	$c = 50$	4.24	3.47	3.04	2.75	2.55	2.40	2.28	2.18	2.10
	$c = 100$	4.93	4.17	3.73	3.45	3.25	3.09	2.97	2.88	2.80

For easy reference, Table 3 presents a "ready reckoner" for the γ-index precalculated based on the below-99th citation data[18]. Due to the monotonically increasing property of the cumulative density function, $H(T)$, the γ-index for each citation count monotonically decreases as the elapsed years (T) after being posted on arXiv becomes longer. Note that we have omitted the negative values of γ from the table because, as discussed, the γ-index for the below-99th

[18] The numerics presented in Table 3 are based on the below-99th citation data, and therefore provide a "harsher" evaluation of the citational impact than those numerics calculated with the lower percentile thresholds. For comparison, Supplementary Table S9 presents another ready reckoner for the γ-index precalculated based on the below-90th citation data.

citation data could be conservatively valid for $\gamma \gtrsim 0$. As an illustration, 10 citations at $T = 2$ in the CS discipline ($\gamma = 1.73$) is comparable to 50 citations at $T = 6$ in the same discipline ($\gamma = 1.71$). As another illustration, 10 citations at $T = 3$ in the HEP discipline ($\gamma = 1.68$) is seen as more impactful than the same number of citations at the same elapsed time in the CS discipline ($\gamma = 0.96$). Also, 50 citations at $T = 9$ in the Astrophysics discipline ($\gamma = 2.87$) is comparable to the same number of citations at $T = 4$ in the HEP discipline ($\gamma = 2.91$), or 100 citations at $T = 4$ in the CS discipline ($\gamma = 2.85$).

We have demonstrated how the idea of the quantitatively and temporally normalized index of the citational impact works by applying the γ-index to the real world (arXiv) data. However, it should be kept in mind that this index would still produce a significantly biased result if applied without an appropriate classification of research disciplines or subfields. For instance, even within the same discipline of HEP, papers on theoretical HEP and experimental HEP could show quite distinct characteristics regarding the typical citation history profile. The same situation can also occur for the Math discipline (e.g., between papers on pure mathematics and applied mathematics). We also note that, depending on the research discipline, the γ-index may not be a beneficial indicator of the citational impact for particularly "young" papers produced within a couple of years (cf. Wang, 2013). With all these caveats, the γ-index (or the γ^*-index) can offer a useful, albeit approximate and imperfect, measure to evaluate and compare the citational attention on scientific publications across disciplines and time periods.

4.2. Stochastic Modeling of Fundamental Citation Dynamics

So far, we have investigated the macroscopic, discipline-average picture of citation distribution and evolution pattern. Indeed, the regression model of Eq. 5 achieved a remarkable fit to the empirical arXiv data. However, the underlying microscopic mechanism that governs the quantitative and temporal dimensions of citation dynamics at the level of individual papers, consistent with the revealed macroscopic picture, is yet to be uncovered. To address this issue, we provide mathematical modeling of the citation dynamics through a stochastic differential equation (SDE). A key observation is that the lognormal law observed for the degree distribution is related to geometric Brownian motion. This stochastic model plays a central role in many mathematical models, most notably in Black–Scholes option pricing (Black & Scholes, 1973; Merton, 1973) known in the financial engineering literature. Recall that the Black–Scholes model involves the deterministic and stochastic parts, reminding us of some structural similarity to the citation dynamics. In the real world of scholarly communication via citations, when an author of scientific papers determines which literature to cite, the author's attitudes and behaviors are often influenced—whether consciously or unconsciously—by the recent citation performance of the literature. However, that does not solely determine the paper's reference list; the author's finite cognitive capacity, resource limitations, time constraints, and pursuit of novelty introduce some random, nondeterministic factors in the citation phenomenon. Other random events would include the discovery of unexpected relationships to other works and the subsequent development of research. These observations lead us to speculate that the citation dynamics is also described by a deterministic plus stochastic model, as in the Black–Scholes model. Then, intuition suggests that the deterministic part is encoded in the average citation history function, whereas the stochastic part introduces the statistical variability in the trajectory distribution of citations. It will be an extended Black–Scholes model, where the "drift" and "volatility" parameters are now time dependent, encoding the nontrivial temporal characteristics of the citation dynamics. Below, we show that it is indeed the case.

Let $\Delta c(t_i, t_{i+1})$ be the number of citations a paper receives during the time interval $[t_i, t_{i+1}]$ with $t_0 = 0$ and $\Delta t = t_{i+1} - t_i$, $i \in \mathbb{Z}_0^+$. We interpret $\Delta c(t_i, t_{i+1})$ as a discretized realization of a continuous latent attention function, $X(t)$, evaluated at t_i, such that $\Delta c(t_i, t_{i+1}) = \lfloor X(t_i)\Delta t \rfloor$, where the floor function is defined by $\lfloor x \rfloor := \max\{n \in \mathbb{Z} | n \le x\}$ for $x \in \mathbb{R}$. The latent attention function, $X(t)$, is assumed to be strictly positive for all $t \in \mathbb{R}_0^+$. The cumulative citation count at time $t = T = N\Delta t$, denoted as $c(T)$, is related to $X(t)$ as $c(T) = \sum_{i=0}^{N-1} \Delta c(t_i, t_{i+1}) = \sum_{i=0}^{N-1} \lfloor X(t_i)\Delta t \rfloor \approx \int_0^T X(t)\, dt$ for sufficiently small Δt, or large N[19]. Then, the problem of finding the underlying SDE boils down to obtaining $X(t)$ satisfying the following four properties:

1. $X(t)$ is continuous and positive for for all $t \in \mathbb{R}_0^+$.
2. $X(t)$ consists of a deterministic part and a stochastic part.
3. The mean value of $X(t)$ reproduces the observed average citation history curve.
4. The distribution of cumulative citations for a large ensemble of papers follows the observed lognormal law.

Skipping all the details of the derivation (see Appendix C in the Supplementary Materials), the SDE whose solution satisfies Properties 1–4 is given by

$$dX(t) = X(t)[d\ln u(t) + \beta(t)dW(t)]. \tag{14}$$

Here, $\{W(t) | t \in \mathbb{R}_0^+\}$ is the standard Brownian motion (or the Wiener process) with $W(t + \Delta t) - W(t) \sim \mathcal{N}(0, \Delta t)$, $u(t) \in \mathbb{R}^+$ is the average citation history function as before, and $\beta(t) \in \mathbb{R}^+$ controls the random fluctuations of $X(t)$ during the time interval dt as a response to external random events discussed above. The SDE (Eq. 14) can be solved straightforwardly, in the Itô's sense, yielding

$$X(t) = u(t)\exp\left(-\frac{1}{2}\int_0^t \beta(\tau)^2 d\tau + \int_0^t \beta(\tau)dW(\tau)\right) = \frac{u(t)}{w(t)}\exp\left(\int_0^t \beta(\tau)dW(\tau)\right), \tag{15}$$

where we introduced $w(t) := \exp(\frac{1}{2}\int_0^t \beta(\tau)^2\, d\tau)$. The probability density function for $X(t)$ is obtained by solving the corresponding Fokker–Planck equation. Here we only present the final result (see Appendix C in the Supplementary Materials for the derivation):

$$p(x, t) = \frac{1}{x\sqrt{4\pi \ln w(t)}}\exp\left[-\frac{\ln^2(xw(t)/u(t))}{4\ln w(t)}\right]. \tag{16}$$

We can check that Eqs. 15 and 16 readily satisfy Properties 1–4. Property 1 is automatically satisfied with a continuous positive function $u(t)$. Property 2 is already ensured at the level of the SDE (Eq. 14), which is also manifest in the solution (Eq. 15). Property 3 can be directly checked by calculating the expectation value on both sides of Eq. 15, yielding $\mathrm{E}[X(t)] = u(t)$. Consequently, if we choose the citation history function specifically to be $u(t; \hat{\Omega}) = \hat{A}f(t + 1; \hat{\mu}, \hat{\sigma}) + \hat{B}g(t; \hat{\lambda})$ with f and g defined in Eqs. 6 and 7, respectively, then the solution (Eq. 15) precisely reproduces the evolution pattern observed for the arXiv citation data. Finally, regarding Property 4, the probability density function (Eq. 16) reproduces the lognormal distribution Eq. 3 under the identification of $\hat{b} \sim \ln(u(T)/w(T))$ and $\hat{m} \sim \sqrt{2\ln w(T)}$. Via the latter identification, $w(t)$ can be constructed through the information extracted from $\hat{m}$ (Supplementary Table S8). Also, $u(t)$ can be obtained from the estimated regression parameters presented in Table 2. With the so obtained $w(t)$ and $u(t)$, the resulting probability density function (Eq. 16) is shown to exhibit an extremely skewed lognormal distribution at each time section, where most of the probability density is concentrated in the region $x \ll 1$. Therefore, the distribution of $c(T) \approx \int_0^T X(t)\, dt$ is also approximated by a lognormal

[19] More precisely, the relation $c(T) \le \lfloor \int_0^T X(t)dt \rfloor \le c(T) + 1$ follows from the property of the floor function.

distribution, reproducing the empirical lognormal law; thus, Property 4 is satisfied. It is noteworthy that, as is clear from the above construction, the proposed stochastic model of citation dynamics is not only applicable to the typical jump–decay plus constant attention citation pattern but also various atypical patterns (He et al., 2018; Ke et al., 2015; van Raan, 2004). Moreover, it would also be applicable to a wide range of studies concerning popularity and collective attention (Lorenz-Spreen, Mønsted et al., 2019; Simon, 1971).

5. SUMMARY AND CONCLUSIONS

This study investigated the citations of more than 1.5 million eprints on arXiv to model and analyze the quantitative and temporal characteristics of collective attention on scientific knowledge. The eprints included not only papers published in journals but also their preprint versions and other open electronic materials. The developed conceptual and technical framework allowed us to explore the citation dynamics beyond the traditional journal-centered scholarly communication, making the previously invisible visible. By applying a nonlinear regression model to the long-term discipline-average citation data, some interesting facts were revealed regarding how quickly and how many, on average, an eprint in each discipline acquires citations and how quickly it becomes obsolete. Apart from very highly cited papers, it was uncovered that while the jump–decay plus constant attention patterns were consistent across disciplines, the quantitative characteristics such as the average peak height, the average time to reach the peak, and the curve's skewness varied among the disciplines. In particular, CS exhibited the steepest gradient with the highest peak in the growth phase, while Math exhibited the heaviest-tailed curve profile in the obsolescence phase. The regression results were used to analyze the trends in the obsolescence rate, in both the quantitative and temporal dimensions. The temporal obsolescence rate, or the internal obsolescence rate, was quantified by $\mathcal{S} = \delta_1/\delta_2$, where δ_1 and δ_2 represented the typical time interval of the growth phase and the obsolescence phase, respectively, identified in the average citation history curve. The quantity-wise obsolescence rate, or the retention rate, was quantified by $\mathcal{R} = \hat{B}/\hat{u}_\mathrm{p}$, where $\hat{B}$ and $\hat{u}_\mathrm{p}$ represented the asymptotic value and the peak value of the average citation history curve, respectively. Both obsolescence metrics exhibited varied characteristics across disciplines, reflecting the different publication and citation practices. Further, the revealed characteristics of the discipline-average citation history curves were used to develop a new citation index, called the γ-index, which was normalized across disciplines and time periods. When applied to the arXiv citation data, the distribution of γ was fairly close to the standard normal distribution (except its left tail). This fact suggested that the γ-index could be used as an improved, less biased alternative to those citation-based impact measures widely used in the current academic or government practice. Moreover, a stochastic modeling of the citation dynamics is presented, which successfully reproduced both the observed lognormal law for the cumulative citation distribution—the quantitative dimension—and the observed time behavior of the average citation history curve—the temporal dimension—in a unified formalism.

The new conceptual and methodological framework developed in this paper to explore the dynamics of collective attention on science would be of interest to a wide range of research communities, including academic researchers, practitioners, and policymakers. However, as with any bibliometric research, this study also faced various limitations that may have impacted the general validity of the findings. First, papers not being posted on arXiv, including those directly submitted to journals, were not considered due to the data availability. Although a growing number of research outputs in the fields of physical, mathematical, and computer sciences have relied on arXiv (Supplementary Figures S1a, b), those citations outside the

current arXiv data could have impacted some quantitative aspects of the findings. Second, the results should be assessed with caution, as they are likely to be highly dependent on the research discipline classification scheme (Schubert & Braun, 1996; Waltman, 2016; Zitt et al., 2005). Although this study relied on the same classification as that used on the arXiv.org (2021) website, different specifications based on different subject areas or specialities could also have been applied, leading to quantitatively different implications on the citation dynamics. The validity and utility of the γ-index also wholly depends on the discipline classification scheme. Third, the citation data would have been incomplete due to the limitation of the external source and its connectivity to the arXiv system. There have not been consistent tools to measure citations in the sphere of eprints, and arXiv's citation-tracking algorithm may also be imperfect. If a significant number of papers or related citations have not been correctly recorded in the Semantic Scholar API data, then the quantitative results could have been different from—albeit qualitatively similar to—the findings of this paper. Fourth, and finally, as also emphasized in the discussion on the γ-index, it should be noted that there are inherent limitations in using citation-based methods to evaluate the research impact. Citation is not necessarily a measure of the absolute quality or value of scientific works; its primary criterion is utility in research (Garfield, 1979). Therefore, our quantitative approaches must not solely be used to evaluate individual research outcomes and must always be combined with expert judgements or trustworthy peer reviews.

There are many directions in which this work may be extended or applied. For instance, the quantitative information on the discipline-average citation history curve can be used to adjust the citation window to calculate various popular citation-based metrics, including the journal impact factor (Garfield, 1972, 2006) and the *h*-index (Hirsch, 2005), alleviating the problem of biased assessment. Such knowledge will also be helpful to better detect "atypical" citation patterns for individual papers, including the cases of multiple humps and Sleeping Beauties (He et al., 2018; Ke et al., 2015; van Raan, 2004). It would also be interesting to investigate the determinants of various bibliometric indicators considered in this paper, including the internal obsolescence rate, the retention/inflation rate, and the γ-index (γ*-index). Our compiled arXiv data set contained the information of the arXiv ID, which can be used to identify various information about each eprint, including title, author names, affiliations and the number of authors, institutions to which the authors belong, researcher ID, journal titles, and DOI (if applicable). The researcher ID can also be used to draw information about the authors' research records and various achievements. By integrating these information resources into a more comprehensive data set, we will be able to conduct causal inference with both quantitative and qualitative methods for a deeper understanding of the citation dynamics. In so doing, it is important to extend the research area beyond the arXiv disciplines, including the areas of not only other natural sciences and biomedical sciences but also humanities and social sciences. The yet-to-be-seen landscape of interdisciplinary research driven by open eprints would be different from that seen through the lens of journal papers (e.g., Okamura, 2019). Finally, it would be interesting to investigate further how the networks of citations (Barabási et al., 2002; Golosovsky & Solomon, 2017; Pan et al., 2018; Wang et al., 2013), or more broadly, collective attention (Lorenz-Spreen et al., 2019) is formed and disseminated through time and social space, based on the framework developed in this paper. The increasing availability of large-scale data will enable us to explore new frontiers in scientometrics.

We close this paper with a remark on what "citation" would represent in the coming new era of Open Science. As ever more scientific knowledge is posted, published and made available to the public through various open digital materials and platforms, those who "cite" the source of knowledge will be diversified across the stakeholders of science and society.

Simultaneously, the reasons and ways for citation will also be diversified. The scientific landscape and the scenery, which researchers have long seen over the shoulder of giants, will necessarily change as Open Science provides a new way to see the world from high above the data clouds. It will enable people to access and experience collaborative knowledge production and consumption processes through open, transparent and shared platforms. Accordingly, what the term "citation"—or scientific attention, attitudes, and behavior—has meant and means today in scientific communities will not be the same tomorrow. A deeper understanding of the *who*, *what*, *why*, and *how* of "citation" in a much broader context will be required to build a balanced and sustainable relationship between science and the future society.

ACKNOWLEDGMENTS

The author would like to thank Ryo Suzuki and two anonymous reviewers for their valuable comments on the manuscript. The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of any of the organizations to which the author is affiliated.

COMPETING INTERESTS

The author has no competing interests.

FUNDING INFORMATION

The author did not receive any funding for this research.

DATA AVAILABILITY

The data sets generated and/or analyzed during this study can be found in the Zenodo repository at https://doi.org/10.5281/zenodo.5803962.

REFERENCES

Abdill, R. J., & Blekhman, R. (2019). Tracking the popularity and outcomes of all bioRxiv preprints. *eLife*, *8*, e45133. https://doi.org/10.7554/eLife.45133, PubMed: 31017570

Aksnes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open*, *9*(1), 1–17. https://doi.org/10.1177/2158244019829575

Aman, V. (2013). The potential of preprints to accelerate scholarly communication: A bibliometric analysis based on selected journals. *Masters Thesis*, School of Library and Information Science, Humboldt University of Berlin.

arXiv.org. (2021). *arXiv submission rate statistics*. Data for 1991 through 2020, updated 1 January 2021. https://arxiv.org/help/stats/2020_by_area/.

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512. https://doi.org/10.1126/science.286.5439.509, PubMed: 10521342

Barabási, A.-L., Jeong, H., Néda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, *311*, 590–614. https://doi.org/10.1016/S0378-4371(02)00736-7

Barnett, G. A., Fink, E. L., & Debus, M. B. (1989). A mathematical model of academic citation age. *Communication Research*, *16*(4), 510–531. https://doi.org/10.1177/009365089016004003

Berg, J. M., Bhalla, N., Bourne, P. E., Chalfie, M., Drubin, D. G., ... Wolberger, C. (2016). Preprints for the life sciences. *Science*, *352*(6288), 899–901. https://doi.org/10.1126/science.aaf9133, PubMed: 27199406

Björk, B.-C., & Solomon, D. (2013). The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics*, *7*(4), 914–923. https://doi.org/10.1016/j.joi.2013.09.001

Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy*, *81*(3), 637–654. https://doi.org/10.1086/260062

Bouabid, H. (2011). Revisiting citation aging: A model for citation distribution and life-cycle prediction. *Scientometrics*, *88*(1), 199–211. https://doi.org/10.1007/s11192-011-0370-5

Bouabid, H., & Larivière, V. (2013). The lengthening of papers' life expectancy: A diachronous analysis. *Scientometrics*, *97*(3), 695–717. https://doi.org/10.1007/s11192-013-0995-7

Burrell, Q. L. (2002). The *n*th-citation distribution and obsolescence. *Scientometrics*, *53*(3), 309–323. https://doi.org/10.1023/A:1014816911511

Burton, R. E., & Kebler, R. W. (1960). The "half-life" of some scientific and technical literatures. *American Documentation*, *11*(1), 18–22. https://doi.org/10.1002/asi.5090110105

Eom, Y.-H., & Fortunato, S. (2011). Characterizing and modeling citation dynamics. *PLOS ONE*, *6*(9), e24926. https://doi.org/10.1371/journal.pone.0024926, PubMed: 21966387

Evans, J. A. (2008). Electronic publication and the narrowing of science and scholarship. *Science*, *321*(5887), 395–399. https://doi.org/10.1126/science.1150473, PubMed: 18635800

Feldman, S., Lo, K., & Ammar, W. (2018). Citation count analysis for papers with preprints. *arXiv Preprint*. https://arxiv.org/abs/1805.05238v1

Fraser, N., Brierley, L., Dey, G., Polka, J. K., Pálfy, M., ... Coates, J. A. (2021). The evolving role of preprints in the dissemination of COVID-19 research and their impact on the science communication landscape. *PLOS Biology*, *19*(4), 1–28. https://doi.org/10.1371/journal.pbio.3000959, PubMed: 33798194

Fraser, N., Momeni, F., Mayr, P., & Peters, I. (2020). The relationship between bioRxiv preprints, citations and altmetrics. *Quantitative Science Studies*, *1*(2), 618–638. https://doi.org/10.1162/qss_a_00043

Fu, D. Y., & Hughey, J. J. (2019). Releasing a preprint is associated with more attention and citations for the peer-reviewed article. *eLife*, *8*, e52646. https://doi.org/10.7554/eLife.52646, PubMed: 31808742

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*(4060), 471–479. https://doi.org/10.1126/science.178.4060.471, PubMed: 5079701

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool? *Scientometrics*, *1*, 359–375. https://doi.org/10.1007/BF02019306

Garfield, E. (2006). The history and meaning of the Journal Impact Factor. *Journal of the American Medical Association*, *295*(1), 90–93. https://doi.org/10.1001/jama.295.1.90, PubMed: 16391221

Gentil-Beccot, A., Mele, S., & Brooks, T. C. (2010). Citing and reading behaviours in High-Energy Physics. *Scientometrics*, *84*(2), 345–355. https://doi.org/10.1007/s11192-009-0111-1

Ginsparg, P. (2016). Preprint déjà vu. *The EMBO Journal*, *35*(24), 2620–2625. https://doi.org/10.15252/embj.201695531, PubMed: 27760783

Glänzel, W. (2004). Towards a model for diachronous and synchronous citation analyses. *Scientometrics*, *60*(3), 511–522. https://doi.org/10.1023/B:SCIE.0000034391.06240.2a

Glänzel, W., & Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature. *Journal of Information Science*, *21*(1), 37–53. https://doi.org/10.1177/016555159502100104

Golosovsky, M., & Solomon, S. (2017). Growing complex network of citations of scientific papers: Modeling and measurements. *Physical Review E*, *95*(1), 012324. https://doi.org/10.1103/PhysRevE.95.012324, PubMed: 28208427

Gross, P., & Gross, E. (1927). College libraries and chemical education. *Science*, *66*(1713), 385–389. https://doi.org/10.1126/science.66.1713.385, PubMed: 17782476

Hajra, K. B., & Sen, P. (2005). Aging in citation networks. *Physica A: Statistical Mechanics and its Applications*, *346*(1), 44–48. https://doi.org/10.1016/j.physa.2004.08.048

Hajra, K. B., & Sen, P. (2006). Modelling aging characteristics in citation networks. *Physica A: Statistical Mechanics and its Applications*, *368*(2), 575–582. https://doi.org/10.1016/j.physa.2005.12.044

He, Z., Lei, Z., & Wang, D. (2018). Modeling citation dynamics of "atypical" articles. *Journal of the Association for Information Science and Technology*, *69*(9), 1148–1160. https://doi.org/10.1002/asi.24041

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102, PubMed: 16275915

Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, *112*(24), 7426–7431. https://doi.org/10.1073/pnas.1424329112, PubMed: 26015563

Kim, J. (2019). Author-based analysis of conference versus journal publication in Computer Science. *Journal of the Association for Information Science and Technology*, *70*(1), 71–82. https://doi.org/10.1002/asi.24079

Kirkham, J. J., Penfold, N. C., Murphy, F. L., Boutron, I., Ioannidis, J. P., ... Moher, D. (2020). Systematic examination of preprint platforms for use in the medical and biomedical sciences setting. *BMJ Open*, *10*, e041849. https://doi.org/10.1136/bmjopen-2020-041849, PubMed: 33376175

Larivière, V., Archambault, E., & Gingras, Y. (2008). Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004). *Journal of the American Society for Information Science and Technology*, *59*(2), 288–296. https://doi.org/10.1002/asi.20744

Larivière, V., Sugimoto, C., Macaluso, B., Milojević, S., Cronin, B., & Thelwall, M. (2014). arXive-prints and the journal of record: An analysis of roles and relationships. *Journal of the Association for Information Science and Technology*, *65*(6), 1157–1169. https://doi.org/10.1002/asi.23044

Lehmann, S., & Lautrup, B. (2006). Measure for measures. *Nature*, *444*(7122), 1003–1004. https://doi.org/10.1038/4441003a, PubMed: 17183295

Line, M. B. (1993). Changes in the use of literature with time: Obsolescence revisited. *Library Trends*, *41*(4), 665–683.

Line, M. B., & Sandison, A. (1974). "Obsolescence" and changes in the use of literature with time. *Journal of Documentation*, *30*(3), 283–350. https://doi.org/10.1108/eb026583

Lorenz-Spreen, P., Mønsted, B. M., Hövel, P., & Lehmann, S. (2019). Accelerating dynamics of collective attention. *Nature Communications*, *10*(1), 1759. https://doi.org/10.1038/s41467-019-09311-w, PubMed: 30988286

Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado López-Cózar, E. (2016). Back to the past: On the shoulders of an academic search engine giant. *Scientometrics*, *107*(3), 1477–1487. https://doi.org/10.1007/s11192-016-1917-2

Merton, R. C. (1973). The theory of rational option pricing. *The Bell Journal of Economics and Management Science*, *4*(1), 141–183. https://doi.org/10.2307/3003143

Merton, R. K. (1968). The Matthew effect in science. *Science*, *159*(3810), 56–63. https://doi.org/10.1126/science.159.3810.56, PubMed: 5634379

Moed, H. F. (2007). The effect of "Open Access" on citation impact: An analysis of arXiv's Condensed Matter section. *Journal of the American Society for Information Science and Technology*, *58*(13), 2047–2054. https://doi.org/10.1002/asi.20663

Nakamoto, H. (1988). Synchronous and diachronous citation distributions. *Informetrics*, *87–88*, 157–163.

Newman, M. E. J. (2001). Clustering and preferential attachment in growing networks. *Physical Review E*, *64*, 025102. https://doi.org/10.1103/PhysRevE.64.025102, PubMed: 11497639

Nicolaisen, J. (2007). Citation analysis. *Annual Review of Information Science and Technology*, *41*(1), 609–641. https://doi.org/10.1002/aris.2007.1440410120

Okamura, K. (2019). Interdisciplinarity revisited: Evidence for research impact and dynamism. *Palgrave Communications*, *5*(141), 1–9. https://doi.org/10.1057/s41599-019-0352-4

Okamura, K., Yoda, H., Hayashi, K., & Koshiba, H. (2020). *Current issues around preprints and the implications for science and technology policymaking*. A collaborative policy paper (in Japanese) by Ministry of Education, Culture, Sports, Science and

Technology (MEXT) and National Institute of Science and Technology Policy (NISTEP), Japan.
Okamura, K., & Koshiba, H. (2021). Citation data of arXiv eprints and the associated quantitatively-and-temporally normalised impact metrics ('γ-index') [Data set]. *Zenodo*. https://doi.org/10.5281/zenodo.5803962
Pan, R. K., Petersen, A. M., Pammolli, F., & Fortunato, S. (2018). The memory of science: Inflation, myopia, and the knowledge network. *Journal of Informetrics*, *12*(3), 656–678. https://doi.org/10.1016/j.joi.2018.06.005
Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K., & Fortunato, S. (2015). Attention decay in science. *Journal of Informetrics*, *9*(4), 734–745. https://doi.org/10.1016/j.joi.2015.07.006
Peterson, G. J., Pressé, S., & Dill, K. A. (2010). Nonuniversal power law scaling in the probability distribution of scientific citations. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(37), 16023–16027. https://doi.org/10.1073/pnas.1010757107, PubMed: 20805513
Price, D. J. d. S. (1965). Networks of scientific papers. *Science*, *149*(3683), 510–515. https://doi.org/10.1126/science.149.3683.510, PubMed: 14325149
Price, D. J. d. S. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, *27*(5), 292–306. https://doi.org/10.1002/asi.4630270505
Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Towards an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, *105*(45), 17268–17272. https://doi.org/10.1073/pnas.0806977105, PubMed: 18978030
Radicchi, F., Weissman, A., & Bollen, J. (2017). Quantifying perceived impact of scientific publications. *Journal of Informetrics*, *11*(3), 704–712. https://doi.org/10.1016/j.joi.2017.05.010
Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B – Condensed Matter and Complex Systems*, *4*(2), 131–134. https://doi.org/10.1007/s100510050359
Redner, S. (2005). Citation statistics from 110 years of *Physical Review*. *Physics Today*, *58*(6), 49–54. https://doi.org/10.1063/1.1996475
Schubert, A., & Braun, T. (1996). Cross-field normalization of scientometric indicators. *Scientometrics*, *36*(3), 311–324. https://doi.org/10.1007/BF02129597
Seglen, P. O. (1998). Citation rates and journal impact factors are not suitable for evaluation of research. *Acta Orthopaedica Scandinavica*, *69*(3), 224–229. https://doi.org/10.3109/17453679809000920, PubMed: 9703393
Sevryugina, Y. V., & Dicks, A. J. (2021). Publication practices during the COVID-19 pandemic: Biomedical preprints and peer-reviewed literature. *bioRxiv Preprint*. https://doi.org/10.1101/2021.01.21.427563
Sheridan, P., & Onodera, T. (2018). A preferential attachment paradox: How preferential attachment combines with growth to produce networks with log-normal in-degree distributions. *Scientific Reports*, *8*(2811), 1–11. https://doi.org/10.1038/s41598-018-21133-2, PubMed: 29434232
Shuai, X., Pepe, A., & Bollen, J. (2012). How the scientific community reacts to newly submitted preprints: Article downloads, Twitter mentions, and citations. *PLOS ONE*, *7*(11), e47523. https://doi.org/10.1371/journal.pone.0047523, PubMed: 23133597
Simon, H. A. (1971). Designing organizations for an information-rich world. In M. Greenberger (Ed.), *Computers, communications, and the public interest*. Johns Hopkins Press.
Sinatra, R., Deville, P., Szell, M., Wang, D., & Barabási, A. (2015). A century of physics. *Nature Physics*, *11*(10), 791–796. https://doi.org/10.1038/nphys3494
Thelwall, M. (2016). Are the discretised lognormal and hooked power law distributions plausible for citation data? *Journal of Informetrics*, *10*(2), 454–470. https://doi.org/10.1016/j.joi.2016.03.001
Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do altmetrics work? Twitter and ten other social web services. *PLOS ONE*, *8*(5), e64841. https://doi.org/10.1371/journal.pone.0064841, PubMed: 23724101
van Raan, A. F. J. (2004). Sleeping Beauties in science. *Scientometrics*, *59*(3), 467–472. https://doi.org/10.1023/B:SCIE.0000018543.82441.f1
Verstak, A., Acharya, A., Suzuki, H., Henderson, S., Iakhiaev, M., ... Shetty, N. (2014). On the shoulders of giants: The growing impact of older articles. *arXiv Preprint*. https://arxiv.org/abs/1411.0275v1
Waltman, L. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, *10*(2), 365–391. https://doi.org/10.1016/j.joi.2016.02.007
Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127–132. https://doi.org/10.1126/science.1237825, PubMed: 24092745
Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, *94*(3), 851–872. https://doi.org/10.1007/s11192-012-0775-9
Wang, Z., Chen, Y., & Glänzel, W. (2020). Preprints as accelerator of scholarly communication: An empirical analysis in Mathematics. *Journal of Informetrics*, *14*(4), 101097. https://doi.org/10.1016/j.joi.2020.101097
Yin, Y., & Wang, D. (2017). The time dimension of science: Connecting the past to the future. *Journal of Informetrics*, *11*(2), 608–621. https://doi.org/10.1016/j.joi.2017.04.002
Zhang, L., & Glänzel, W. (2017). A citation-based cross-disciplinary study on literature aging: Part I—The synchronous approach. *Scientometrics*, *111*(3), 1573–1589. https://doi.org/10.1007/s11192-017-2289-y
Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics*, *63*(2), 373–401. https://doi.org/10.1007/s11192-005-0218-y