Science of science—Citation models and research evaluation[*)]

Vincent Traag 0000-0003-3170-3879[†)]
Centre for Science and Technology Studies (CWTS), Leiden University, the Netherlands

(Dated: 12th May 2025)

Citations in science are being studied from several perspectives, among which approaches such as scientometrics and science of science. In this chapter I briefly review some of the literature on citations, citation distributions and models of citations. These citations feature prominently in another part of the literature which is dealing with research evaluation and the role of metrics and indicators in that process. Here I briefly review part of the discussion in research evaluation. This also touches on the subject of how citations relate to peer review. Finally, I conclude by trying to integrate the two literatures. The fundamental problem in research evaluation is that research quality is unobservable. This has consequences for conclusions that we can draw from quantitative studies of citations and citation models. The term "indicators" is a relevant concept in this context, which I try to clarify. Causality is important for properly understanding indicators, especially when indicators are used in practice: when we *act* on indicators, we enter causal territory. Even when an indicator might have been valid, through its very use, the consequences of its use may invalidate it. By combining citation models with proper causal reasoning and acknowledging the fundamental problem about unobservable research quality, we may hope to make progress.

Keywords: science of science; citations; research evaluation; peer review

The study of science itself has a venerable history, and is studied from several points of view. The field of scientometrics studies science from a quantitative perspective. Relatedly, the field of science of science is similarly taking a quantitative perspective, but often with a somewhat different approach. The two have much in common and share a more quantitative formal perspective on studying science, especially based on large-scale data sets of publications, their authors and their citations. Scientometrics has been traditionally more focused on "measuring science", while much of science of science is more focused on "modelling science". This distinction is not absolute though: some publications in what most would consider scientometrics build models, while publications in science of science sometimes also address issues of measuring. I will review part of this literature, with a focus on citations.

Studies are often motivated by the fact that citations are considered to be relevant for how science operates: they may reflect advances in science and clarify intellectual contributions. Moreover, citations and related aspects seem to play a role in scientists' own careers, a development that seems to have become increasingly stronger over the years. The use of metrics in research evaluation is increasingly criticised. The role of metrics in research evaluation and the effects of metrics are less often discussed explicitly by the scientometric and science of science literature. In this chapter I aim to connect these two literatures, with a focus on citations.

First I discuss various observations of citation distributions, how they change over time, and how they can be modelled. This literature is largely based on a mix of scientometrics and science of science. Then, I review a small part of the literature on research evaluation. This includes some aspects relevant to national research evaluations. This also touches upon issues of comparing peer review and metrics, which I will also briefly discuss. After reviewing this literature, I will conclude by bringing the two literatures in conversation with each other. I deconstruct some aspects of citation dynamics and clarify that other factors play a role in citation dynamics, the implications of which are, although sometimes acknowledged, not often appreciated in this literature. Additionally, I will consider how we can think of "indicators" in this context, how they can be biased, and how they can be made more accurate. A causal understanding is key to understanding indicators, I believe.

The fundamental problem in research evaluation is that scientific quality is unobservable. Any study on the subject therefore must acknowledge this, and this has consequences for the type of conclusions that we can draw, especially from quantitative studies. Being more clear about our causal reasoning and being careful about what we can and cannot conclude helps to clarify that. Having a better understanding of the overall dynamics in citations, and relying on models that capture these dynamics, we can improve what we can infer from observations.

[*)] This is a draft. The final version will be available in *Handbook of Computational Social Science* edited by Taha Yasseri, forthcoming 2025, Edward Elgar Publishing Ltd. The material cannot be used for any other purpose without further permission of the publisher and is for private use only. Please cite as: Traag, V. A. (2025). Citation models and research evaluation. In: T. Yasseri (Ed.), *Handbook of Computational Social Science*. Edward Elgar Publishing Ltd.

[†)] Electronic mail: v.a.traag@cwts.leidenuniv.nl

I. CITATIONS

A. Citation distributions

One of the earliest and most commonly studied aspects in scientometrics and science of science is the distribution of the number of citations. Various authors have tried to find theoretical distributions that could fit the empirical distribution well. Part of the literature has tried to come up with theoretical models that could explain the observed type of distributions, and I will cover some such studies later. One important consideration here is that there does not exist such a thing as *the* distribution of citations. Distribution of citations always refer to a particular set of papers, and the results will vary across fields, years, journals or institutions.

One of the earliest studies of citations was covered by Price (1965). He studied the number of citations to all papers covered in one of the earliest editions of the Science Citation Index, the precursor of what is currently known as the Web of Science. Price (1965) finds that citations are distributed approximately as a power law:

$$\Pr(C \geq c) \propto c^{-\alpha+1}, \tag{1}$$

with α estimated to be somewhere between 2.5–3.0. This points to a highly skewed citation distribution. Indeed, he finds that “only 1 percent of the cited papers are cited as many as six or more times each in a year”.

Physicists became increasingly interested in citations and citation networks in the late 1990s. Redner (1998) is an early study of citations in that literature and reports a power law distribution with an exponent of about 3, similar to Price (1965). He studies a few different citation distributions: a distribution from a single year (1981) coming from a precursor of the Web of Science, and a few different volumes of *Physical Review D*. These different datasets show quite a different number of average citations and older years generally have accumulated more citations, simply as the result of having had more time to accumulate citations. He finds that various datasets collapse onto a universal curve when dividing the number of citations by the average number of citations in that dataset.

Laherrère and Sornette (1998) study a slightly different citation distribution, namely the citation distribution of all citations to authors, instead of citations to individual papers. They find that a stretched exponential is the best fit for their distribution:

$$\Pr(C \geq c) = \exp - \left(\frac{c}{c_0}\right)^\alpha \tag{2}$$

However, they also find that a power law is a reasonable fit, with an exponent of about 3 again.

Radicchi, Fortunato, and Castellano (2008) study citation distributions of a few different fields and a few different years. They study these distributions separately and try to determine whether the distributions are universal, in the sense that after some transformation, all the distributions look alike. They find that a simple scaling of citations with the average number of citations in the same field and the same year collapses all the distributions onto a single curve, hence finding evidence for universal scaling. That is, they define the normalised citations $\tilde{C}_i = \frac{C_i}{\mathrm{E}(C_i)}$ where C_i is the total number of citations received for publication i and $\mathrm{E}(C_i)$ is the average number of citations received over all publications from the same field and the same publication year. They find that the normalised citations $\tilde{C}_i$ are well-fitted by a lognormal distribution $\mathrm{LogNormal}(-\frac{\sigma^2}{2}, \sigma^2)$, with $\sigma^2 \approx 1.3$, which by definition has an average of 1. In this study they used data from Web of Science and relied on the field definitions given by journal subject categories. They repeat this study later with data from the American Physics Society (APS) and PACS codes to define fields (Radicchi and Castellano, 2011), again finding a similar collapse of distributions onto a universal curve. Chatterjee, Ghosh, and Chakrabarti (2016) perform a similar study of the universality of citation distributions, but then of institutions and journals. They also find evidence for universality and find that the normalised citations are well-fitted by a lognormal distribution. For academic institutions they find that $\sigma^2 \approx 1.7$; somewhat more skewed than the paper level citation distribution identified by Radicchi, Fortunato, and Castellano (2008), while for journals the collapsed citation distributions is slightly less broad with $\sigma^2 \approx 1.4$. For both institutions and journals, the lognormal is less able to fit the tail of the distributions, which seems to be better approximated by a power law. Possibly, this could be a result of differences in sizes, which plays a role for institutions and journals, but not for individual paper distributions.

The universality claim by Radicchi, Fortunato, and Castellano (2008) was revisited by Waltman, Van Eck, and Van Raan (2011), who argued that citation distributions are not truly universal, and that differences can still be observed between some fields. They study this by comparing the top 10% of all publications based on the normalised citations to the top 10% within each field. If citations were perfectly universal, the overall top 10% would overlap with the top 10% of each field, but this is not the case. Ignoring uncited articles does make the case for universal distribution stronger. The probability of having zero citations may be slightly distinct from the overall citation distribution.

In a follow-up analysis Radicchi and Castellano (2012) devise a clever way of empirically deriving a slightly different normalisation such that citation distributions across different fields collapse. They base this on comparing the overall citation distribution to all distributions of citations per field, and find that the transformation $\left(\frac{C}{a}\right)^\alpha$ produces highly similar distributions across nearly all fields of science, where a and α are estimated empirically. For this universal distribution, they find it is well fit by a lognormal distribution.

As suggested by the analysis of Waltman, Van Eck, and Van Raan (2011), the case of zero citations may function slightly differently. Wallace, Larivière, and Gingras (2009) also consider these uncited articles when studying a century of citation distributions. In particular, they find that $e^{\beta \frac{N_r}{N_a}}$ is a good fit for predicting the number of uncited papers in a distribution, where N_a is the total number of articles published in a year and N_r the total number of references to those N_a articles. This is based on a simple idea that the N_r citations are randomly distributed across the N_a articles, and the uncitedness is the probability of having drawn 0 references, at least within a short time window (2 years). They fit a stretched exponential to the citation distribution, where the probability to be cited c times is

$$\Pr(C) \sim P(0) \exp - \left(\frac{C}{\tau} \right)^{\alpha} \tag{3}$$

where τ and α are estimated parameters, with $P(0)$ the separately modelled uncited publications. This is based on the idea that different papers accumulate citations at different rates, and that the overall distribution is a mixture of all those different rates. It is not clarified how the stretched exponential arises as a mixture of individual Poisson processes with different rates. One possibility is to model the distribution as a mixture of Poisson distributions with the rate of each Poisson distribution following a Gamma distribution. That would result in a Negative Binomial distribution, which is studied by Mingers and Burrell (2006). Thelwall and Wilson (2014) find that Negative Binomial regression is a bad fit, and advise against using it, and suggest using an OLS logarithmic fit. To cover the entire range, Wallace, Larivière, and Gingras (2009) suggest that a distribution first suggested by Tsallis and de Albuquerque (2000) fits best:

$$\Pr(c) = \frac{P(0)}{[1 + (q-1)\lambda c]^{\frac{q}{q-1}}}, \tag{4}$$

with parameters λ and q, but a clear theoretical underpinning for this distribution is lacking.

The decline of concentration in citations is described by Larivière, Gingras, and Archambault (2009). They find that over time, the number of uncited papers continues to decrease (except for the humanities). Whereas in the 1920s about 70% of the articles remain uncited within 5 years, in the 2000s this has decreased to about 10-30%. The citation distribution also seems to become less skewed over time. Before the Second World War, the percentage of papers that attracted 80% of the citations increased from a few percent to 25–30%, and it continues to hover around that percentage, with most recent times seeing an even larger increase.

Redner (2005) also takes a long-term perspective, studying citation statistics from 110 years of *Physical Review* journals. He finds that the overall number of citations of all papers is well fit by a lognormal distribution. This is in a sense surprising, since he studies the distribution of papers from multiple years (1893–2003), in which case you might expect a mixture of yearly lognormal distributions, which could result in a stronger power law tail.

Moreira, Zeng, and Amaral (2015) find that a (discretised) lognormal distribution captures well citation distribution over sets of papers from authors and departments, and that the distributions are relatively stable over time. Sinatra *et al.* (2016) find that a lognormal distribution also fits well the citation distribution over a set of papers. Stringer, Sales-Pardo, and Amaral (2005) find that a lognormal distribution also fits well the citation distribution over journals, and that the distribution becomes stationary after about 10 years. Similar to Milojević, Radicchi, and Bar-Ilan (2016), they use this to rank journals by focusing on the probability that a paper from one journal is cited more highly than a paper from another journal. The ranking results are consistent with journal distributions being approximately lognormal.

Overall then, the most reasonable assumption seems to be that citations are distributed approximately as a lognormal. Other observed distributions most likely arise as mixtures of a lognormal, resulting in stronger power law tails.

B. Temporal decay of citation rate

Citations generally decay over time. Most papers tend to cite recent work more frequently than older work. We can study this from two perspectives. We can take a retrospective, backward looking approach (Burrel, 2001), sometimes called a synchronous approach (Line and Sandison, 1974), and study the age of references in papers. Alternatively, we can take a prospective, forward-looking approach (Burrel, 2001), sometimes called a diachronous approach (Line and Sandison, 1974), and study at how frequently a paper is cited in the years after it is published.

The decay of citations over time is sometimes referred to as obsolescence, referring to the decline of the use of certain publications over time. Publications need not become fully obsolete, but their usage may decline nonetheless. As Line and Sandison (1974) explain, there are various reasons why certain publications may become obsolete. The work may become common knowledge in the field, sometimes referred to as obliteration by incorporation (Garfield, 1957). This happens for example when a theory has become eponymised, such as the Nash equilibrium (McCain, 2011). Alternatively, the work may have become outdated or belong to an abandoned paradigm (Kuhn, 2012). Work may also later be found to be incorrect or inaccurate (Furman, Jensen, and Murray, 2012), although some citations continue after retractions for example, seemingly unaware of the retracted status (Bornemann-Cimenti, Szilagyi, and Sandner-Kiesling, 2016).

Many studies take a retrospective approach. This approach is easier to use, especially historically. Taking a retrospective approach involves taking a paper, and checking the years of the cited references in that paper, which is relatively straightforward, especially when working with actual printed papers, which used to be the case historically. In contrast, as Egghe and Rousseau (2000) point out, a prospective approach requires one to go through all papers to check whether it has cited the paper of interest, This is only possible if there is a proper citation database available.

Gross and Gross (1927) are one of the first to study how publications reference literature in earlier years. Burton and Kebler (1960) introduced the half-life of attention/usage in this context, although half-life was already used earlier, according to Line (1970), while the concepts of growth, utility and obsolescence were introduced by Brookes (1970). Price (1965) introduced a measure of immediacy, later sometimes called the Price Index, which is defined as the percentage of references younger than t years.

Line (1970)[1] distinguishes between real and apparent obsolescence, arguing that we should control for the number of papers being published, which increases exponentially each year. In his words “if every item had an equal probability of being used or cited, more use would be made of more recent literature simply because there is more of it.” He takes a retrospective approach and studies the (median) reference age. He introduces a very simple correction to the observed obsolescence factor. Suppose that the observed obsolescence is $a(t) = \frac{c(t-1)}{c(t)}$, where $c(t)$ is the total citations given to articles in the year t, from some reference year $t' > t$. Now suppose that the number of citations $c(t)$ has grown from year $t-1$ to t with a factor $g(t)$ such that $c(t) = g(t)c(t-1)$. In order to correct the obsolescence $a(t)$ for this growth $g(t)$, we should then divide $c(t-1)$ by the expected number of citations $\frac{c(t)}{g(t)}$ that were obtained had there been no growth. Hence, the growth corrected obsolescence should then be defined as $a(t) = \frac{c(t-1)}{\frac{c(t)}{g(t)}}$. Assuming constant growth rates, $a(t) = a$ and $g(t) = g$, we then obtain constant corrected obsolescence $d(t) = d$. The growth-corrected number of citations in year t then simply is $c(t) = c(t-1)d$, such that $c(t) = c(0)d^t$, and the infinite series $\sum_t c(t)$ equals $\frac{c_0}{1-d}$. The corrected half-life h is then $\frac{\log \frac{1}{2}}{\log a + \log g}$ while the uncorrected (observed) half-life would be $\frac{\log \frac{1}{2}}{\log a}$. Although a gross oversimplification, it nicely captures how the growth in the number of publications affects the apparent obsolescence of the literature. With a yearly growth percentage of 5%, a median citation age of 7 years would suggest that items might be considered for removal from the library after 7 years, while in reality they would continue to be used for almost 14 years.

Brookes (1970) discusses some problems with estimating the obsolescence, and relates this to geometric decay of utility such that $c(t) \propto (1-a)a^{t-1}$ with an annual ageing factor a. Egghe and Ravichandra rao (1992) argue against the ageing perspective from Brookes (1970) that assumes a constant ageing factor. Instead, they find that ageing has a certain minimum, suggesting there is a natural peak in reference age. They find that the most sensible distribution is then a lognormal distribution, based on finding a unique minimum in ageing, and find it fits the data well.

Avramescu (1979) studies retrospective reference distributions. He suggested the following model to fit to the retrospective distribution:

$$c(t) = C_0 \left[\exp(-\alpha t) - \exp(-m\alpha t)\right] \tag{5}$$

where $c(t)$ is the number of citations received t years after publication, and α and $m > 1$ are some parameters.

Instead of working with obsolescence rates $a(t) = \frac{f(t-1)}{f(t)}$ Egghe (1994) proposes a continuous counterpart for a continuous function $c(t)$, namely $a(t) = \exp(\log f(t))'$, where the prime $'$ indicates taking the derivative with respect to t. This of course equals $\exp\left(\frac{f'(t)}{f(t)}\right)$ so that this is the exponent of the relative growth of t. This is a “true” rate, as Egghe (1994) states, and makes intuitive sense and has some sensible properties. However, this formulation does not seem to have been used frequently.

Stinson and Lancaster (1987) studies citation ageing from both a synchronous and diachronous perspective. They take into account a correction for the growth of the literature, but they do not report any particular distribution.

Redner (2005) finds an exponential decrease in the age distribution. As suggested by Nakamoto (1988), the growth in the number of publications is also relevant in this context. Combining the exponential decrease in age of references with an exponential growth of publication leads to a power law decrease in age overall (Redner, 2005; Egghe, 2005).

Vinkler (1996) formulates a relatively simple model for the possibility to be cited and finds that the possibility to be cited increases with the growth of the field. Faster growing fields are hence more likely to show higher chances of citations. This is also noticed by Hargens and Felmlee (1984) who argue that in growing fields, older work tends to gather more citations from recent work than in stable fields. Scientists might therefore be eager to jump on the bandwagon of a newly emerging field, because it pays off to be one of the first movers in a new field.

Larivière, Archambault, and Gingras (2008) take a long-term perspective, and find that the average reference

[1] Interestingly, this seems to be one of the earliest examples of a paper that appends the report by one of the referees, at least that I am aware of. An early example of transparent peer review, even signed by the referee!

age increased over the last decades. Similarly, in physics the average reference age was found to have increased over the last decades (Sinatra *et al.*, 2015). Verstak *et al.* (2014) also find that the average reference age increased over the last decades in various fields. Larivière, Archambault, and Gingras (2008) observe some interesting peaks during both world wars. Relatively few publications were published during those two periods, showing a dip in the number of publications. Most papers that appeared during, and shortly after, the war therefore reference papers from before the war, resulting in a quite high reference age.

Egghe (2010) proposes a simple model for some observations of increasing reference age, as observed by Larivière, Archambault, and Gingras (2008), while it still has a decreasing Price Index (i.e. the proportion of references in the last t years). The model is quite straightforward: it assumes that the literature grows exponentially, and that publications are cited completely at random. Even in that simplest case, one already sees an increasing reference age, but a decreasing Price Index. Hence, qualitatively, such observations do not require an explanation beyond a simple exponential growth of the literature.

Parolo *et al.* (2015) study the prospective citation distribution and state that the nature of the decay is not well established, varying between an exponential decay and a slower power law decay. They find that attention decays faster more recently than in earlier years. If they renormalise time in terms of the number of papers, they find that the decay rate is stable. Hence, the faster attention decay is simply a result of the increasing number of publications. They only study the decay after the initial peak of citations. Over time, the peak in citations has come increasingly faster, consistent with the increasing reference age found by Larivière, Archambault, and Gingras (2008), according to Parolo *et al.* (2015). The decay after the initial peak is best fit by an exponential function. The half-life decreases over time, and citations taper off increasingly faster in more recent years. Again, when rescaling time in terms of number of publications, this decrease is no longer visible.

Šubelj and Fiala (2017) find that the peak year of reference distributions (i.e. retrospective) has stayed stable in computer science and physics. The peak year of the citation distribution (i.e. prospective) has shifted however, and is more volatile, especially for computer science. Again, when normalising the citations based on the number of publications, all distributions seem to collapse onto a universal curve. As Egghe and Rousseau (2000) explain, growth influences ageing, but it does not cause ageing per se. They find that increasing growth rates lead to higher obsolescence, i.e. papers tend to become obsolete more quickly.

Pan *et al.* (2018) also study the ageing of reference distributions, and finds evidence of “citation inflation”: papers need increasingly more citations to be part of the top 5%. They find that citations to recent literature and very old literature decreased, while citations to the “middle” part increased.

Gingras *et al.* (2008) find that the average age of the references depends on the age of researchers. Younger researchers initially tend to cite more recent work, but when researchers become older, their references age with them, with a turning point when researchers become about 40 years old.

Herman (2004a,b) studies scholars’ literature search behaviour qualitatively. She finds that most people only go back a few years to look for references to keep up-to-date on the most recent development. Most scholars mentioned that they would not search the literature further back than just a couple of years, but may follow up by chasing down references from that literature.

Poncela-Casasnovas *et al.* (2019) find that papers that reference a highly cited paper and are relatively highly cited themselves as well are published relatively shortly after each other. This suggests something like the start of a field, where an initial publication is cited by another paper shortly afterwards, both of which play a role in the ensuing citation dynamics and the influx of authors to such a field. Higher impact papers tend to cite younger papers and very young papers (< 1 year). They find that method references are typically older. Bertin *et al.* (2016) find that references in the introduction of a paper are typically older. This most likely sets the stage and background of a field for a paper.

In principle, there is a certain connection between a retrospective and a prospective approach. The exact connection depends on the dynamics of the number of publications and the number of references throughout time. But, in general, if the retrospective distribution remains stable throughout time, it can be used to infer the prospective distribution, while using the empirically observed publication and referencing dynamics. Yin and Wang (2017) provide an exact relationship between the two approaches and find that

$$\overleftarrow{\Pr}(t_2 \mid t_1)M(t_1) = \overrightarrow{\Pr}(t_1 \mid t_2)L(t_2) \tag{6}$$

with $\overleftarrow{\Pr}$ the prospective distribution and $\overrightarrow{\Pr}$ the retrospective distribution, where $M(t) = m(t)N(t)$ is the total number of references given at time t, with $m(t)$ the average number of references in year t and $N(t)$ the total number of publications in year t, which approximately grows exponentially over time $N(t) \sim e^{\beta t}$, while $L(t) = \int_t^\infty \overleftarrow{\Pr}(t \mid \tau)M(\tau)d\tau$ is the total number of citations received by papers at time t. Hence, one can derive the one distribution from the other. Arguably, the retrospective distribution is primary and the prospective distribution is derivative. After all, the retrospective distribution describes how researchers behave and choose to cite previous literature, while the prospective distribution is the result of that process.

Yin and Wang (2017) find that after normalising the citations for the number of publications it is well fit by a lognormal distribution (both prospective and retrospective). This entails that the unnormalised, crude, age distributions are a mixture of the publication and referencing dynamics and the actual lognormal decay.

C. Citation models

The models that I cover in this section attempt to capture various observations. Some models try to explain the overall citation distribution, others target the ageing distribution of references, while others aim to model individual paper citation dynamics, sometimes with an eye on predicting future citations.

One of the first models that was introduced in this context was developed by Price (1976). It introduced the notion of cumulative advantage, based on the ideas of the Matthew effect introduced earlier by Merton (1968), sometimes called a rich-get-richer effect. The model of Price (1976) aims to explain the broad distribution of citations observed earlier (see also section I A). The model is relatively straightforward and works as follows. For each time step, an additional paper is added to the population, citing m earlier papers. These references are not added randomly, but are assumed to be distributed proportional to the current number of citations of each publication. That is, the probability of paper i to be cited is proportional to $C_i(t)$, the total number of citations at time step t. Often some constant is added, in order to make sure that papers for which $C_i(t) = 0$ also have some non-zero probability to be cited. The overall citation distribution is then affected by the influx of new papers at each time step, which initially have no citations, and the earlier publications which accumulate increasingly more citations. These forces give rise to a distribution, which Price (1976) calls the Cumulative Advantage Distribution $c \sim (m+1)B(c, m+2)$ where $B(a, b)$ is the Beta function. In the limit of large citations c this approaches a power law with exponent $m + 2$, which for $m = 1$ is close to the earlier observations of citation distributions reviewed in section I A. This idea of cumulative advantage was again suggested in the late 1990s in the context of complex networks by Barabási and Albert (1999), who termed this preferential attachment.

Redner (2005) finds some evidence for a linear preferential attachment, and suggests that a redirection mechanism could be reasonable. That is, instead of directly connecting to a paper with probability proportional to $C_i(t)$, the idea is to pick a reference from a randomly selected paper (with probability $1 - r$) or simply reference the randomly selected paper itself (with probability r), leading to a linear preferential attachment.

Demonstrating that there is a cumulative advantage effect in empirical observations is not easy. Often scholars study the relationship between the cumulative number of citations $C(t)$ after time t and the additional citations in some time period Δt after t. If the cumulative number of citations $C(t)$ is correlated with this increase $\Delta C(t+\Delta t) = C(t+\Delta t) - C(t)$, this is often taken as evidence for the existence of a cumulative advantage. However, this does not need to be the case. The inherent problem with this approach is that some *latent citation rate* of the article may affect both $C(t)$ and $\Delta C(t + \Delta t)$. Hence, the additional citations $\Delta C(t + \Delta t)$ need not be the result of the earlier citations $C(t)$: publications that achieve a higher $C(t)$ just have a higher latent citation rate, and therefore also show a higher $\Delta C(t + \Delta t)$. As a straightforward example, if citations accrue at a rate of λ then $E(C(t)) = \lambda t$ and so $\mathrm{E}(\Delta C(t+\Delta t)) = \lambda \Delta t$, so that $\mathrm{E}(C(t)) = \mathrm{E}(\Delta C(t + \Delta t))\frac{t}{\Delta t}$. Hence, observing a linear growth of the additional number of citations with the initial number of citations need not indicate much more than simply the result of a constant growth rate. In other words, the correlation between $C_i(t)$ and $\Delta C(t+\Delta t)$ may not arise due to $C_i(t)$ *causing* more citations $\Delta C(t+\Delta t)$, but because they are confounded by the underlying citation rate λ. Citation distributions might be skewed just because the underlying latent citation rates are skewed, not because of a cumulative advantage effect. In some other contexts there is some clear experimental evidence of a Matthew effect (Van de Rijt *et al.*, 2014). Additionally, there is some qualitative evidence that people browse the literature by following references (Herman, 2004a,b), also leading to a type of cumulative advantage effect. So, there might be other reasons to believe cumulative advantage is reasonable, but the empirical data analysis is more challenging.

Some models start on the basis of such a simple assumption of a constant rate of accumulation of citations, often taking a prospective perspective. Mingers and Burrell (2006) for example proposes a Gamma mixture of Poisson distributions. That is, each paper has a latent citation rate λ_i and the number of citations is then simply Poisson distributed with rate parameter λ_i. Assuming λ_i is distributed according to a Gamma distribution, the overall number of citations is distributed as a Negative Binomial. They fit this distribution to empirical data, and find it to be a good fit, except for some extremely highly cited papers. Next, they suggest that the rate of attracting citations is a time dependent variable $\lambda_i(t)$, with some constant latent citation rate, modulated by some time factor, i.e. $\lambda_i(t) = \lambda_i f(t)$, with $f(t)$ the *obsolescence function*. They estimate their model for journals, estimating the overall citation distribution and the decay, which allows them to predict citations for articles in that journal, but do not predict citations for individual articles. Burrel (2001) considers a similar starting model, where each paper i accumulates citations at a latent citation rate λ_i modulated by some time dependence $f(t)$, so that the effective citation rate is $\lambda_i(t) = \lambda_i f(t)$. In this model, the shape of the distribution of the time to the first-citation is independent of the mixing distribution of the latent citation rates, and only depends on the shape of the obsolescence function, suggesting that the obsolescence function follows some S-shaped pattern. Moreover, after a sufficiently long time, the number of citations depends only on λ_i and not on the obsolescence function $f(t)$. Burrell (2002) follows up on this work and investigates the n-th citation distribution of this model. He finds that a Gamma distribution of latent citation rates, which leads to a Negative Binomial distribution of citations, fits well the data, while relying on an obsolescence function that follows a specific Gamma distribution $\Gamma(2,1)$, corresponding to $f(t) = 1 - e^{-t}(1+t)$.

Higham *et al.* (2017) propose that the rate of attracting additional citations is a separable function of preferential attachment and some obsolescence function, while taking a forward-looking prospective view. The rate of attracting citations in year t is then

$$\lambda(C(t), t) = a(C(t))f(t), \tag{7}$$

where $f(t)$ is some obsolescence function that depends on time t only and $a(c)$ is some cumulative advantage function that depends on citations $C(t)$ only. In particular, they use functional forms $a(c) = c^\alpha + c_0$, and $f(t) = d_0 \exp\left(-\frac{t}{\tau}\right)$. They test if these can indeed be separated by checking various years and bins of citations c against each other, and find support for the idea of separability. They find that α is about 1.0–1.2 while the exponential fit performs well only for $t \geq 3$. Based on this prospective model, they also derive an expression for the retrospective distribution of references. The separability of citation dynamics is an important observation that simplifies the modelling. Some earlier authors also proposed separable models. Dorogovtsev and Mendes (2000) seems to have introduced the earliest ageing with preferential attachment model, including a separable formulation, and used $f(t) = t^{-\alpha}$ and $a(c) = c$, in terms of Eq. 7. Wang, Yu, and Yu (2009) also proposed a separable model with $f(t) = \exp(-\lambda t)$ and $a(c) = c$, and find it fits well some empirical data. Neither study explicitly addresses the separability though. Hajra and Sen (2006) propose changing earlier models and publish multiple papers simultaneously, instead of sequentially introducing single papers, as was usually done, and find that this improves the fit.

Although the previously discussed models consider obsolescence, they offer no theoretical explanations. Simkin and Roychowdhury (2007) suggest a mathematical theory of citations that provides an explanation for ageing. They propose that every year t there are N papers published that contain N_r references on average. A fraction α of these references goes to randomly selected papers in the preceding year $t-1$ (with $\alpha \approx 0.1$–0.15). This leads naturally to the first-year citations for papers published in $t-1$ being distributed Poisson, in line with earlier discussed results from Wallace, Larivière, and Gingras (2009), with αN_r expected citations. With probability $1-\alpha$ then, a random reference from a random publication in year $t-1$ is followed and is cited. Such a publication from year $t-2$ might have cited a random publication from year $t-2$ (with probability α) or might have cited a random reference from that publication (with probability $1-\alpha$), and so on. This leads to a branching process which Simkin and Roychowdhury (2007) solve analytically. They find the prospective and retrospective distribution of citations to be a power law with an exponential cut-off. To account for large exponential cut-offs and obtain a power law scaling, Simkin and Roychowdhury (2007) propose to add a latent citation rate parameter for each paper. Instead of choosing a random paper and a random reference from a random paper, scientists then choose a paper proportional to the latent citation rate. They consider a uniform distribution of latent citation rates, and marginalise the decay over this to obtain a citation distribution across all articles. Various latent citation rate distributions yield similar observed citation distributions.

Peterson, Pressé, and Dill (2010) created a similar model to Simkin and Roychowdhury (2007), but propose as the first step to find random papers from all years instead of the preceding year only. Their model focuses on the citation distribution, not the ageing distribution. Goldberg, Anthony, and Evans (2015) also consider a similar copying model, and find it to be the best fitting model.

Pan *et al.* (2018) propose a model that combines various elements from earlier models. It takes redirection from the models by Simkin and Roychowdhury (2007) and Peterson, Pressé, and Dill (2010), but also uses an initial preferential attachment. In each time step, $n(t)$ new publications are added, which grows exponentially. Each new publication cites directly an existing publication j with probability $(a + C_j(t))f(t_j)^\alpha$ where $C_j(t)$ is the number of citations to publication j, which is published at time $t_j \leq t$, and with an additional k random references from publication j, with k binomially distributed. They find their model to reproduce several stylistic features of citation networks.

Eom and Fortunato (2011) find that a shifted power-law best fits citation distributions. They propose to model the citation network as follows. At each step, a new paper, i.e. node, is added to the citation network. The new paper i cites a previous paper j proportional to their current cumulative number of citations $C_j(t)$ and a certain decay as

$$c_{ij} \propto C_j(t) + \lambda_j f(t) \tag{8}$$

where λ_i is the latent citation rate of article j and $f(t)$ is some decay factor, assumed to be exponential by Eom and Fortunato (2011). They find this model to reproduce various distributions reasonably well.

Wang, Song, and Barabási (2013) introduced a model that similarly combines a temporal decay with a rich-get-richer effect while also allowing for an individual article level parameter to account for variability across papers. In a sense, this approach is similar to what was proposed by Eom and Fortunato (2011), but they only considered aggregate properties, such as citation distributions, whereas Wang, Song, and Barabási (2013) try to predict citation dynamics of individual papers. This hence combines most previous elements, and is relatively similar in spirit to the model by Pan *et al.* (2018). More specifically, Wang, Song, and Barabási (2013) model the rate of attracting additional citations $c_i(t)$ at time t as

$$c_i(t) \propto \lambda_i C_i(t) f(t_i) \tag{9}$$

with $C_i(t)$ the total number of citations up until time t and λ_i the latent citation rate of article i. It might be interesting to empirically compare this model, using a multiplicative formulation, to the earlier model by Eom and Fortunato (2011), which uses an additive formulation. Solving the model by Wang, Song, and Barabási (2013) leads to the result that

$$C_i(t) \propto e^{\lambda_i F(t_i)} - 1 \tag{10}$$

where $F(t_i)$ is the cumulative distribution of $f(t_i)$. For $t_i \to \infty$ then, we have that $F(t_i) = 1$ so that after a sufficiently long time we arrive at

$$c_i(t) \propto e^{\lambda_i} - 1. \tag{11}$$

This means that ultimately, after waiting long enough, the total number of citations is expected to depend only on the latent citation rate λ_i, similar to what was observed by Burrell (2002). Wang, Song, and Barabási (2013) found their model to fit well the citation dynamics of many papers.

In response, Wang, Mei, and Hicks (2014) wrote that the predictions of the model of Wang, Song, and Barabási (2013) were not so good and that naive predictions were more accurate. In a rebuttal Wang *et al.* (2014) argued that overfitting of their model should be prevented by using informative priors (in a Bayesian analysis) or by otherwise regularising the fitting procedure. One element of dispute seems to be the purpose of models. From the perspective of Wang, Mei, and Hicks (2014) the complexity of the model by Wang, Song, and Barabási (2013) is simply not necessary, since a simple prediction performs equally well, while Wang *et al.* (2014) argue that they model the dynamics that are seen in citations. One difference between the two seems to be that, once the model of Wang, Song, and Barabási (2013) is in place, one could in principle predict forward citations across multiple years over time. The naive prediction that Wang, Mei, and Hicks (2014) considered was to actually assume citations after 5 and after 30 years simply have not changed, which is of course not informative. In addition, Penner *et al.* (2013) point to a problem when predicting citations, namely that many studies focus on the cumulative number of citations, which is also relevant in this particular disagreement. Penner *et al.* (2013) argue that comparing cumulative citations is misleading, because one can easily predict cumulative citations from earlier cumulative citations, even if the process is completely random. That is, suppose that $c_i(t)$ is a completely random variable, with the cumulative number of citations up until time t being $C_i(t) = \sum_{\tau=0}^{t} c_i(\tau)$. The correlation of $C_i(t)$ between time t and $t + \Delta t$ can then be quite high and equals

$$\sqrt{\frac{t}{t + \Delta t}}. \tag{12}$$

Hence, if Δt is small compared to t, the correlation will be high. For small t, the correlation is lower. These correlations are purely mechanical and result directly from the cumulative citations. If, instead of the cumulative citations $C_i(t)$, we try to predict the yearly citations $c_i(t)$, we would quickly learn that the expected correlation between any $C_i(t)$ and $c_i(t)$ is zero, because the process is completely random. Hence, when comparing the predictive capabilities of different models, the focus should be on predicting $\Delta C(t + \Delta t)$, not on predicting $C(t + \Delta t)$, which, as Wang, Mei, and Hicks (2014) also observe, can nearly trivially be predicted based on $C(t)$.

II. EVALUATION OF RESEARCH

Research is regularly being evaluated, for various reasons, such as funding decisions, hiring decisions or quality assurance. The use and misuse of citation-based metrics regularly feature in the literature on this topic. Here I briefly review some of that literature.

The role of journals in research evaluation has been contested for quite some time. The Journal Impact Factor (JIF)—the average number of citations to a journal in the preceding two years—was originally developed for decisions about journal collection management in libraries (Larivière and Sugimoto, 2019). From the 1990s onwards, JIFs were increasingly used in research evaluation (Hicks *et al.*, 2015). Journals show a high heterogeneity of what they publish, and Seglen (1997) argued that you should not evaluate an individual article based on where it is published, similar to the adage that you should not judge a book by its cover. The JIF became increasingly contested, resulting in a call to abandon them for research evaluation in the Declaration on Research Assessment (DORA, 2013). The subject was also discussed in a workshop on Rethinking JIFs (Wouters *et al.*, 2019). Some even talked about "Impact Factor mania" (Casadevall and Fang, 2014). Following a call to publish the full citation distribution instead of the JIF (Lariviere *et al.*, 2016), the Journal Citation Reports now provide more detailed information. Still, in recent times, the JIF has continued to be used in promotion and tenure decisions (McKiernan *et al.*, 2019).

The JIF was reported to feature not only when evaluating research that has already been done, but also to shape decisions of what research questions to focus on (Rushforth and de Rijcke, 2015). The JIF structures discussions about what is novel and sufficiently high-quality to target high-impact journals. The JIF was not used per se to say something about the potential novelty and quality of the science itself, but was also seen as a "ticket" to advance one's career. Importantly, this shows that the JIF is not just about targeting specific journals once the research itself is already done; the research is done and shaped with impact factors in mind. Indeed, this phenomenon has been called "thinking with indicators", shaping not only post-research where a manuscript should be submitted, or how something is evaluated, but also actively shaping what research is done (Müller and de Rijcke, 2017). These effects of indicator usage have been reviewed more broadly by de Rijcke *et al.* (2016).

One important recurrent theme in this context is that of goal displacement. This phenomenon is sometimes known as Goodhart's law, or Campbell's law: scoring high on assessment indicators becomes more important than doing well on whatever those indicators were meant to measure. This is closely related to the so-called constitutive effects of performance indicators (Dahler-Larsen, 2014). When indicators are used in practice, they may affect how people respond. This should not be thought of as "unintended consequences"; rather, the usage of the indicator itself defines what is evaluated. Hence, if citations are used to evaluate research quality, they might not necessarily be misused, but rather, an indicator, such as citations, comes to represent the very object that they purport to measure. For instance, when publishing university rankings, their very usage alone may result in such rankings becoming thought of as measures of university "performance". More highly ranked universities may attract more students and more high-qualified personnel. Such effects may not necessarily result from university ranking itself being "correct" indicators of performance, but because the ranking itself produces such effects. Something similar may happen with journal impact. Journal impact rankings and publicly visible indicators, such as the JIF, may reify through constitutive effects any initial ranking of "journal impact". That is, if scholars start to judge journals by such a ranking, they might start to submit their best work to the highest ranked journal, which thereby may solidify, or even improve their ranking, while lower ranked journals may start to receive increasingly worse manuscripts, thereby potentially lowering their ranking. In this sense, constitutive effects may function similarly to self-fulfilling prophecies. Whether constitutive effects ameliorate or deteriorate outcomes is not clear *a priori*.

Molas-Gallart and Rafols (2018) provide a broad critique of indicators. Citation-based indicators may not align well with research objectives, leading to an "evaluation gap". They argue that scientists respond to evaluation by aiming to improve their performance as measured by indicators, similar to constitutive effects. If such an evaluation has the desired properties, this effect might be positive, but this need not be the case. Even without responding strategically to such incentives, evaluations may act as a selective pressure (Smaldino and McElreath, 2016). That is, it does not require constitutive effects in order to exert an influence.

Bhattacharya and Packalen (2020) also critique metrics based on the argument that attention (i.e. citations) to novel ideas has decreased, and that evaluating people based on citations effectively selects against novelty. They argue that more scientists are working on only incremental advances that will be more likely to be cited, instead of working on foundational groundwork.

A particular context in which metrics are sometimes used for research evaluation is in performance-based university research funding systems (PBRFS), which were reviewed by Hicks (2012). Although the distribution of funding is an important component of PBRFS, they also seem to feed into a prestige competition. The first and perhaps most well-known PBRFS is the UK's Research Assessment Exercise (RAE), currently known as the Research Excellence Framework (REF). In general, PBRFS aim to stimulate excellence, or fund more selectively, to allocate scarce resources more effectively. The resource concentration has also been linked to the "new public management" that has become more dominant in research policy circles. The most common unit of evaluation is the department of universities or research organisations, although some countries also evaluate individual scientists, for example for appointing professors.

There is an extensive literature discussing the potential effects of PBRFS. Butler (2003) performed a seminal study on the increase of publications in lower impact journals following the introduction of a PBRFS in Australia. Another effect of introducing PBRFS is that researchers may cite each other's work more heavily (Baccini, De Nicolao, and Petrovich, 2019). Some authors found evidence that self-citations increased after the introduction of an evaluation system for promotion in Italy (Seeber *et al.*, 2017). Moed (2008) showed that the UK RAE exercises seemed to affect UK scholars' publishing practices. The classical work by Butler (2003) was revisited by Van den Besselaar, Heyman, and Sandström (2017), reaching different conclusions: productivity and impact both increased in the Australian case. However, generally, causes and effects in PBRFS are rather challenging to disentangle (Aagaard and Schneider, 2017), as argued earlier by Osuna, Cruz-Castro, and Sanz-Menéndez (2011). Gläser and Laudel (2016) describe the overall problem of inferring how macro level science policies affect macro level outcomes. Their central question is: How does research governance change knowledge production? This not only needs to be studied at the macro level, which is bound to be affected by problematic confounding effects (e.g. other changes happening simultaneously); this also needs to be studied at a micro level, providing evidence for a macro-micro-macro link. That is, it should be made reasonable that the macro policy affects researchers' behaviour, which in turn becomes visible at the system level again. One additional potential problem is that an increase in national productivity may also increase national citations. Such higher within-country citations are regularly observed (Schubert and Glänzel, 2006; Bakare and Lewison, 2017), similar to citations in the same language (Bookstein and Yitzhaki, 1999). This raises the question of how to disentangle an increase in citations due to a higher productivity from an increase in citations due to actual differences in research quality. Whether such observations are really driven by national citation biases, or whether they are a result of more general geographical patterns, as observed by Pan, Kaski, and Fortunato (2012) is not clear.

Sandström and Van den Besselaar (2018) study the performance of several national science systems. They conclude that having ex post evaluation, combined with high institutional funding may be most efficient. Ex ante evaluation, either through grant funding, or through lower professional autonomy and more university management, may result in lower efficiency, and may reinforce the existing academic elite.

Schneider, Aagaard, and Bloch (2016) compared the effects of PBRFS in Australia and Norway. They find that, unlike in Australia (Butler, 2003), the introduction of a PBRFS in Norway that awarded publications did not show a decreasing impact or an increasing output in lower impact journals. The important difference here is whether the evaluation differentiates the awards based on some impact indicator. In the Norwegian case they differentiated between lower and higher impact tier outlets. Bloch and Schneider (2016) study the effects of the Norwegian model further, and conclude that due to the fractionalisation, the system may not properly reward collaboration.

In principle, evaluation at the institutional level is to be stimulated (Tiokhin *et al.*, 2021). Institutional evaluation may alleviate some problems that might appear at the individual level, where contributions other than scholarly publications might be disregarded. Institutions can take a broader perspective, and can for example hire someone who does not directly produce scholarly output, but who has a large indirect effect on scholarly output, for instance by maintaining critical infrastructure. Unfortunately, one recurrent problem of evaluation at the institutional level seems to be that institutions pass down the institutional requirements directly to lower levels (Gläser, 2007). For example, in the UK REF system, which is an institutional evaluation, the institutions organise so-called mock-REFs to identify areas where individual scientists could improve their performance, with sometimes dire consequences for their future careers (Owens, 2013).

When studying the causal effects of a PBRFS we should differentiate between system level effects and individual level effects. For example, consider that we fund institutions differentially, based on some performance indicator. After a few years, the overall performance may have increased. At the same time, the differences between institutions may have become smaller: all institutions have increased their performance. Differentiating between institutions then becomes more difficult, and institutions that receive more funding may not necessarily perform much better than institutions that received less funding in one year. It may then appear the differential funding may not be predictive of individual performance, while at the same time, the differential funding did increase the overall quality.

A. Peer review

Most scientists argue that the scientific “quality” of a paper is a multidimensional concept (Aksnes, Langfeldt, and Wouters, 2019). For example, in most journals peer review is based on multiple criteria, such as novelty, potential impact and methodological rigour. In recent years, peer review has been heavily discussed, with multiple possible interventions on several fronts, such as open peer review, post-publication peer review or collaborative peer review (Woods *et al.*, 2022). In almost any evaluative setting, the focus is on trying to evaluate research “quality”. The question is how either peer review or citations can reflect such “quality”. Let me briefly review some of the literature on peer review.

Bornmann (2011) provides a general overview of peer review and identifies a number of problems of peer review. One particular problem is poor reliability: the inter-rater reliability between peer reviewers is generally low. This was already observed earlier by Cole, Cole, and Simon (1981), but was also confirmed in later research again by Ernst, Saradeth, and Resch (1993), Rothwell and Martyn (2000) and Pier *et al.* (2018). The low reliability of peer review opens up the possibility of bias. When a decision needs to be made in a difficult case, the possibility for bias becomes larger to “tip the scale”. On the other hand, the uncertainty in peer review can be one of its strengths. It is difficult to know in advance how something will be evaluated by peers, so using peer review for evaluation decreases the chances of people targeting a specific indicator. Low agreement on evaluation may also reflect different positions and considerations that reviewers may have on a manuscript. Peer review can indeed improve the reporting of findings (Goodman *et al.*, 1994), although the textual changes are often relatively minor (Klein *et al.*, 2016). In a sense, poor agreement demonstrates that multiple reviewers provide more comprehensive feedback than a single reviewer. If reviewers would simply reiterate the same point, there is little added value of the additional reviewer. Initiatives, such as the consultative peer review from eLife (King, 2017), try to benefit from this diversity and suggest an innovative approach to consolidate the various points raised by multiple reviewers.

As said, one problem of peer review is the potential bias: factors unrelated to “quality” may affect peer review (Lee *et al.*, 2013). It can be challenging to establish whether something is a bias (Traag and Waltman, 2022). For example, simply showing that authors from a particular institution have higher peer review scores is insufficient: it is possible that such authors simply more often produce higher quality work. Comparisons of double-blind to single-blind peer review reveal some interesting effects, where author and affiliation reputation seem to affect the acceptance of manuscripts (Tomkins, Zhang, and Heavlin, 2017; Okike *et al.*, 2016).

Another problem that Bornmann (2011) identifies is that of validity: peer review might be unable to predict scientific impact or relevance. However, the problem is that scientific impact and peer review itself may be noisy: how will we measure scientific impact? For example, if you compare the "best" unfunded scholars (i.e. those with the highest scientific impact) to the scientific impact of funded scholars, as done by Van den Besselaar and Sandström (2015), it might very well be that the "best" unfunded scholars outperform the funded ones, not because the best unfunded are "better" than the funded, but simply because citations are such a noisy proxy (Lai, Traag, and Waltman, 2020). Bornmann and Daniel (2008b) analyse the citation outcomes of both accepted and rejected publications at the prestigious *Angewandte Chemie International Edition*, and find that peer review outcomes predict subsequent citations. However, this conclusion is problematic if the publication venue causally affects how frequently it is cited (Traag, 2021). In that case, citations do not necessarily reflect whether peer review is predictive, they may just reflect the causal effect of being published in a certain venue. A similar problem plays in an analysis of the predictive validity of peer review when highlighting publications in a journal (Antonoyiannakis, 2021).

B. Metrics

Much research in scientometrics does not necessarily engage with the citation models that I briefly covered in section I. Instead, much research is interested in factors that somehow seem to affect citations, ranging from the effects of authors to institutions. Some also study aspects such as title length, number of pages and other characteristics, but I will ignore those studies here. Most of the more quantitative studies do not explicitly use any citation model, but simply compare different articles with each other in one way or another, and try to draw conclusions from that comparison. Other studies focus on the meaning of citations, and study the different types of "influence" that citations capture.

MacRoberts and MacRoberts (1989) provide a comprehensive overview of some of the problems in citation analysis. Although their overview is over thirty years old by now, many of the identified problems are still playing a role, and continue to be discussed and studied. I have already covered some common problems in section I, namely varying citation patterns in different fields, years and document types. Another category of problems concerns whether citations really capture the idea of "influence" or impact: not all influences are cited and some works are cited that have no influence (so-called perfunctory citations). There are different types of citations (Bornmann and Daniel, 2008a), which do not show an equal influence, with some citations for example being negative (Lamers *et al.*, 2021). This does not necessarily mean that highly-cited publications are not influential. For example, Teplitskiy *et al.* (2020) find that highly-cited publications are actually more likely to have an intellectual influence on the work in which they are cited. Another category of problems mentioned by MacRoberts and MacRoberts (1989) is more technical and relates to coverage issues (Visser, Van Eck, and Waltman, 2021), problems of reference matching (Olensky, Schmidt, and Van Eck, 2016) and problems of author disambiguation (Caron and Van Eck, 2014).

Co-authored papers are cited more frequently, and this holds for multiple authors, multiple institutions and multiple countries (Larivière *et al.*, 2015). This seems not a result of self-citation, but really represents greater "epistemic value", as stated by Larivière *et al.* (2015). Wu, Wang, and Evans (2019) have looked at this from a slightly different angle and found that larger teams typically produce less disruptive papers, but they are more likely to be more highly cited.

Cole and Cole (1968) find that the prestige of a department affects the visibility of authors. Cole (1970) finds that the prestige of a department also affects (early) citation counts, especially for work that is of lower quality. Similarly, Medoff (2006) finds that institutional prestige drives citations in economics, but only for elite universities. Way *et al.* (2019) find evidence that research quality is driven by scholars' current work environment, and that it is not driven by selection of more highly cited scholar into more prestigious departments.

The role of journals in citations has been debated for a long time. As I already discussed earlier, citation distributions of journals are roughly lognormal. Correlations between the JIF of a journal, and the individual citations for each article is generally low (Seglen, 1997). A more recent revisit of the work by Seglen (1997) again found that correlations between impact factors and citations are relatively low (Zhang, Rousseau, and Sivertsen, 2017). It was also shown that the correlation between the JIF and citations has weakened over the years (Lozano, Larivière, and Gingras, 2012), which was speculated to have been caused by digitalisation. Electronic publication was observed to narrow the referencing, also to more recent literature (Evans, 2008). At the same time, where an article is published is one of the strongest single predictive factor of citations in several studies (Stegehuis, Litvak, and Waltman, 2015; Callaham, 2002; Abramo, D'Angelo, and Di Costa, 2010; Mingers and Xu, 2010).

As already stated earlier, the fundamental problem is that research quality is unobservable. Clearly, citation distributions are highly skewed for each journal, and also overlap to a large extent, as I discussed earlier. However, citations are only a proxy of quality, and are not equal to research quality. Similarly, being published in a certain journal may be a proxy of quality. The question is then: which is a better proxy? Although many people may argue that citations are a more accurate proxy, this need not be the case, as Waltman and Traag (2020) demonstrate. It is possible that all articles within the same journal have the same quality and that the broad distribution of citations is simply due to citations being a noisy proxy of this identical quality. The average of these noisy citations can then be a more accurate representation of the underlying identical quality than the actual citations. The extent to which journals publish similar quality articles is up for debate. This for example will depend on reviewer uncertainty when scholars submit publications. If there is substantial uncertainty, and reviewers try to assess the actual quality of the papers, then the resulting distributions of quality in journals may largely overlap (Starbuck, 2005).

High-impact journals are more widely circulated, and hence have a higher readership (Peritz, 1995). There is a certain circularity here, and path dependency: higher impact journals have a higher readership, which attracts more interesting submissions, which in turn attracts more readers, which in turn attracts more citations. Ellis and Durden (1991) found that current journal prestige is mostly determined by previous journal prestige and current impact, lending some support to this idea of path dependency and conservatism of journal prestige.

More generally, publicity has clear effects on citations. Phillips *et al.* (1991) analysed what papers were being discussed in the New York Times, and how that influenced citations ten years later. Using a three-month period during which the NYT did not appear, but the editorial process and selection remained, they studied the causal effect of publicity in the NYT. They found a quite strong effect: featured papers received 73% more citations. At the same time, the newsworthiness itself also predicts the impact of the journal in which an article will appear (Callaham, 2002).

Citations to identical papers showed that versions that were published in more highly cited journals were cited more often (Knothe, 2006; Perneger, 2010), which was also coined as the Impact Factor's Matthew effect (Larivière and Gingras, 2010). Seglen (1994) questioned whether there was any causal relation between JIF and citations. I will get back to this in section III.

C. Comparing peer review and metrics

Metrics have been regularly compared to peer review outcomes. Both are thought to be indicators of scientific "quality" or "impact", and both have been used in research evaluation. One central difference is that metrics can only be used post-publication, while peer review is also used frequently pre-publication, for example when reviewing journal submissions. Many national PBRFS I discussed earlier, such as the UK REF, the Italian VQR or the Norwegian system are post-publication evaluation systems, and some are based explicitly on metrics (such as the Norwegian model), while others are based on peer review (such as the UK REF) or a mixture of the two (such as the Italian VQR). In the influential Metric Tide report (Wilsdon *et al.*, 2015), the use of metrics in the national research evaluation in the UK was extensively discussed. They concluded that metrics could support but not supplant peer review, as also summarised by Wilsdon (2015).

Aksnes and Taxt (2004) compare peer review and metrics in Norway. They find that normalised citations correlate best with peer review evaluations at the research group level, and report higher correlations for higher aggregate levels. The average journal impact shows a similar level of correlation with peer review. Interestingly, when considering citations relative to the journal (i.e. controlling for the journal impact) they find the lowest correlation.

Bornmann and Leydesdorff (2013) find that peer review, in the form of recommendations from F1000, is correlated with a number of citation-based indicators. Noticeably, this again finds that when normalising based on the journal, there is barely any correlation between peer review and the journal-normalised citation-based indicator.

Radicchi, Weissman, and Bollen (2017) asked respondents to compare pairs of papers, and asked them which paper had a higher influence on their own work. Generally, they find a rather low correlation between citations and those pairwise preferences, but for respondents' own papers, more highly cited papers were more often said to have a higher influence on their own work.

Adams, Gurney, and Jackson (2008) compared evaluation outcomes of papers in the RAE with journal-normalised citation scores, and found that they essentially did not correlate. Eyre-Walker and Stoletzki (2013) also showed that correlations between evaluation and citations are minimal when controlling for the journal, although some of their conclusions have been questioned by Eisen *et al.* (2013).

There are a few problems when comparing peer review and metrics. Studies have reported wildly varying correlations, ranging from as low as 0.3 to as high as 0.97. Traag and Waltman (2019) argue that the comparison between peer review and metrics should take into consideration at least two aspects:

1. Whether size-dependent (e.g. a sum) or size-independent (e.g. a mean) indicators are compared.

2. What level of aggregation (e.g. individual papers, departments, entire universities) is being analysed.

Depending on these choices, correlations can be high or low. In addition, Traag and Waltman (2019) argue that correlations should also be compared to a "baseline" of peer review uncertainty.

Analysis of data from the Italian VQR exercise shows that peer review is not very reliable (Bertocchi *et al.*, 2015), as I already discussed earlier. Compared to correlations between two peer reviewers, correlations between peer review and metrics are found to be comparable. This holds not only at the individual paper level (Bertocchi *et al.*, 2015) but also at the aggregate institutional level (Traag, Malgarini, and Sarlo, 2020). The correlations at the institutional level are typically higher, and this holds both for correlations between two peer reviewers and between peer review and metrics. Of course, an average evaluation outcome can be estimated more accurately when using more peer reviewers. Hence, in the case of many peer reviewers, repeating an evaluation exercise should give highly similar answers, but may still leave a difference between peer review and metrics. However, the resources to do such a large scale peer review exercise are generally limited, so this is infeasible in practice. A recent study by Forscher *et al.* (2019) reported that in the context of NIH funding, one would need as many as 12 reviewers to obtain a modest reliability in funding decisions.

III. CONCLUSION

In this chapter, I briefly reviewed both citation models and some relevant aspects of research evaluation, including peer review and metrics. Although they are treated separately, citation models and research evaluation are related, and the two literatures can be brought into closer conversation with each other.

First, citation models may help us draw inferences about certain effects. As I reviewed in section II B, there are many questions about how various factors may or may not affect citations, such as author reputation, institutional reputation and journal reputation. However, the inference of these effects is tricky. Models such as the one by Wang, Song, and Barabási (2013) may help to disentangle such effects. Essentially, the rate at which an article attracts citations can be formulated as $\lambda(t)$, where $\lambda(t)$ can be composed of multiple various factors, such as authorship, affiliation status, nationality, language, or the journal. Following Wang, Song, and Barabási (2013), the number of citations at time t can then in general be modelled as

$$c(t) \sim \lambda(t) f(t) C(t) \tag{13}$$

with $C(t) = \sum_{\tau=0}^{t} c(\tau)$ the cumulative number of citations. As said, $\lambda(t)$ can be composed of various factors, and we could for instance consider $\lambda(t) = \prod_k \phi_k(t)$ as a product of factors $\phi_k(t)$, which may include factors like author reputation, affiliation reputation, journal reputation, novelty, interdisciplinarity, methodological rigour, data quality, et cetera. In general, this formulation would be highly degenerate: the overall rate λ may be caused by a higher ϕ_1 or a higher ϕ_2 and it is not clear how we can properly identify and estimate the effects of these various factors separately. With additional assumptions, some of these effects may sometimes be estimated. For example, one can consider differences in citation rates when authors become affiliated with other institutions, as was done by Way *et al.* (2019), or one can compare preprints to their journal publications, as was done by Traag (2021).

Secondly, citation models provide more clarity about uncertainty. They clarify that, even for a single paper, the number of citations is not uniquely pre-determined. That is, for each observed outcome, a different outcome might have been observed, if the entire citation dynamics had been replayed. Even considering a simple Poisson process, there is quite some variation in the realised citations, and so inferring the latent citation rate based on the observed citations will show quite some uncertainty. Other factors, such as cumulative advantage process may increase the uncertainty even further.

It is good to explicitly consider uncertainty. For example, for an early career researcher, we might observe only a few papers and a few citations. When using empirical means, or other aggregate statistics, we might easily reach overly extreme conclusions when ignoring the uncertainty. Explicitly considering the uncertainty in such statistics, for example using a Bayesian approach with informed priors, might provide much more reasonable estimates of performance, shrinking the observed number of citations towards more reasonable estimates. As another example, Antonoyiannakis (2018) argues that smaller journals tend to have more extreme citation averages as a result of the law of large numbers: smaller samples will show more variation. Explicitly modelling the uncertainty might help provide more reasonable estimates of journal performance. Similar arguments could be made for estimates of citation impact of research groups, departments, or entire institutions.

Third, when building citation models, we should acknowledge the fundamental problem: *research quality is unobservable.* This means we cannot simply rely on citation models to draw inferences of research quality or "academic success". However, citation models can help clarify how citations could potentially serve as an indicator for research quality. Let us develop a preliminary notion of what an indicator is. We could define an indicator as any variable that is causally affected by the variable for which it serves as an indicator. So, if $X \to \ldots \to Y$, then Y is an indicator for X, with the arrows representing a causal effect. Typically, we do not know X and we therefore use Y to say something about X, and it is in this sense that Y is an indicator for X. However, what is typically the case is that some other factor U also affects Y. In this case, Y might still be an indicator for X, but if U is not exclusively affected by X, we could say that Y is a *biased* indicator for X. After all, we use Y to say something about X in this context, but Y is also affected by U, which is not relevant for saying something about X. For example, if we have Q the "quality" of an article and C citations, where it is assumed that $Q \to \ldots \to C$, then citations are an indicator for quality. However, if citations C are also influenced by another factor U that is deemed irrelevant, such as author affiliation, then using citations C as an indicator for quality Q would be biased by the affiliation U.

We usually (implicitly) assume that the quality Q is unrelated to some other factors that are related to citations, in particular the field F and year Y. That is, we usually assume that $F \to C$ and $Y \to C$, but that F and Y are independent of Q otherwise (Fig. 1). Under the as-

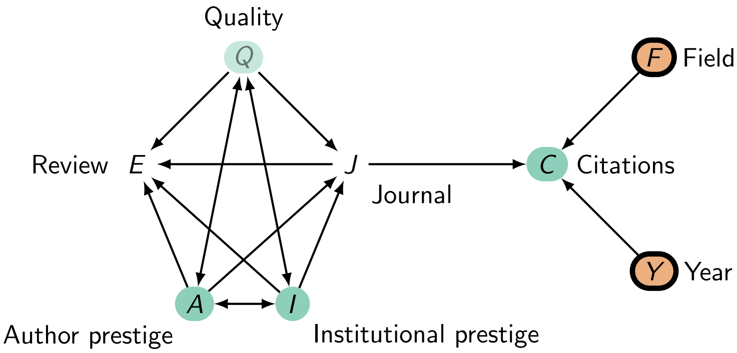

Figure 1. Possible causal model of how citations C can act as an indicator for research quality Q. Here the field F and the year Y are assumed to be independent of quality Q, so that normalising citations C by field F and year Y improves the accuracy of the normalised indicator for quality Q.

sumption that F, Y and Q are independent, we can try to make C a more accurate indicator for Q by normalising citations C based on F and Y (Waltman and Van Eck, 2019), which amounts to conditioning on F and Y. Citation normalisation then makes sense from this point of view.

Sometimes, normalisation also considers the document type D, which implicitly assumes that $D \to C$ but that the quality is independent of the document type. However, higher quality work might be more often made available as a research article, instead of for example as an editorial or a letter to the editor. In that case, if we normalise citations by considering the document type, this pathway of quality $Q \to D \to C$ is blocked, and hence, this might actually deteriorate the accuracy of using normalised citations as an indicator for Q.

Normalising citations may make C a more accurate indicator of Q, and if we could observe Q, we could study how the normalisation of citations makes C a better predictor of Q. Directly comparing C to Q is not possible, because of the fundamental problem: research quality Q is unobservable. Instead of comparing C to Q, scholars regularly compare C to another indicator of Q, namely peer review E, as I discussed in section II C. Let us assume that $Q \to \ldots \to E$, which seems a reasonable starting point. By comparing C and E, we hope to learn something about whether C is an accurate indicator of Q. Again, the fundamental problem in research evaluation is that we cannot observe Q, and so any correlation between C and E does not necessarily establish that they are accurate indicators of Q. It merely establishes that C and E are correlated, but this can potentially also be caused by other causal factors. For example, consider that the journal J influences both citations C and peer review E. We would then observe a correlation between C and E, but would be due to the "confounding influence" of journal J, and have nothing to do with Q. In fact, as we discussed, there is empirical evidence that C and E are not correlated after controlling for the journal J. This implies that if citations C are any indicator for quality Q, then only because of the journal J. Indeed, Waltman and Traag (2020) suggested that the journal J might be a more accurate indicator of Q than the citations C.

Now suppose that author prestige A and institutional prestige I affect acceptance for publication in a journal, so that $A \to J$ and $I \to J$, for which there is some evidence, as we saw in section II A. Author and institutional prestige are most likely associated with quality Q (and perhaps mutually reinforce each other). Most likely, A and I also affect the peer evaluation E. However, they cannot directly affect citations C as well, since we have empirical evidence that entire correlation between E and C is due to the journal J.

All in all, this thinking exercise suggests a couple of prestige feedback loops: author prestige, institutional prestige and journal prestige. These prestige cycles are not independent of the underlying quality of the science, and author prestige, institutional prestige, and journal prestige are all related to quality. However, they do seem to obfuscate and confound much of the measurement of research quality by citations, such that citation-based indicators may better be seen as indicators of academic prestige than as scientific impact.

In a sense, these prestige feedback loops may be similar to what O'Neil (2016) referred to as pernicious cycles. By not considering the effects of predictions when people act upon predictions, the predictions themselves become ill-informed, and may potentially have serious consequences. For example, if we use citations to predict institutional scientific performance, scientists may leave certain departments or institutions because of this. On the face of it, citations may then seem to have some predictive value, but it is exactly *because* citations were used to predict scientific performance that resulted in this behaviour. If we had correctly considered the potential effect of citations on this behaviour, we would perhaps have concluded that on the contrary, citations do *not* have any predictive value. This calls attention to clearer considerations of causality (Klebel and Traag, 2024). Whenever an indicator, or a prediction, is used in practice, that is, we *act* on it, we enter causal territory. Even when an indicator initially might have been valid, through its very use, the consequences of its use may invalidate it. This is perhaps what could be called a causal understanding of constitutive effects. By combining citation models with proper causal reasoning and acknowledging the fundamental problem about unobservable research quality, we may hope to make some progress.

IV. FURTHER READING

There are a great number of books and reviews that cover quantitative science studies. An older overview of scientometrics is provided by Hood and Wilson (2001), providing also a history of the origins and various terms related to this field, such as bibliometrics and informetrics. A useful overview of informetrics is provided by Bar-Ilan (2008). De Bellis (2009) provides a comprehensive overview of the field, and also includes some of the more theoretical frameworks that underpin some of the research. Sugimoto and Larivière (2018) cover the essentials of measuring research, and makes for a great introductory read. The science of science approach was briefly reviewed by Fortunato *et al.* (2018), and more recently, was covered in a more accessible form by Wang and Barabási (2021). Some of the literature was also reviewed by Zeng *et al.* (2017) who took a complex network and complex systems approach. A related, but different perspective was offered by Evans and Foster (2011). An overview of some of the literature concerning research evaluation was written by de Rijcke *et al.* (2016).

REFERENCES

Aagaard, K.and Schneider, J. W., "Some considerations about causes and effects in studies of performance-based research funding systems," J. Informetr. **11**, 923–926 (2017).

Abramo, G., D'Angelo, C. A., and Di Costa, F., "Citations versus journal impact factor as proxy of quality: could the latter ever be preferable?" Scientometrics **84**, 821–833 (2010).

Adams, J., Gurney, K., and Jackson, L., "Calibrating the zoom - a test of zitt's hypothesis," Scientometrics **75**, 81–95 (2008).

Aksnes, D. W., Langfeldt, L., and Wouters, P., "Citations, citation indicators, and research quality: An overview of basic concepts and theories:," https://doi.org/10.1177/2158244019829575 **9**, 215824401982957 (2019).

Aksnes, D. W.and Taxt, R. E., "Peer reviews and bibliometric indicators: a comparative study at a norwegian university," Res. Eval. **13**, 33–41 (2004).

Antonoyiannakis, M., "Impact factors and the central limit theorem: Why citation averages are scale dependent," J. Informetr. **12**, 1072–1088 (2018).

Antonoyiannakis, M., "Does publicity in the science press drive citations? a vindication of peer review," (2021), arXiv:2105.08118 [cs.DL].

Avramescu, A., "Actuality and obsolescence of scientific literature," Journal of the American Society for Information Science **30**, 296–303 (1979).

Baccini, A., De Nicolao, G., and Petrovich, E., "Citation gaming induced by bibliometric evaluation: A country-level comparative analysis," PLoS One **14**, e0221212 (2019).

Bakare, V.and Lewison, G., "Country over-citation ratios," Scientometrics **113**, 1199–1207 (2017).

Bar-Ilan, J., "Informetrics at the beginning of the 21st century—a review," J. Informetr. **2**, 1–52 (2008).

Barabási, A.-L.and Albert, R., "Emergence of scaling in random networks," Science **286**, 509–512 (1999).

Bertin, M., Atanassova, I., Gingras, Y., and Larivière, V., "The invariant distribution of references in scientific articles," J. Assoc. Inf. Sci. Technol. **67**, 164–177 (2016).

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., and Peracchi, F., "Bibliometric evaluation vs. informed peer review: Evidence from italy," Res. Policy **44**, 451–466 (2015).

Van den Besselaar, P., Heyman, U., and Sandström, U., "Perverse effects of output-based research funding? butler's australian case revisited," J. Informetr. **11**, 905–918 (2017).

Van den Besselaar, P.and Sandström, U., "Early career grants, performance, and careers: A study on predictive validity of grant decisions," J. Informetr. **9**, 826–838 (2015).

Bhattacharya, J.and Packalen, M., "Stagnation and scientific incentives," Tech. Rep. 26752 (National Bureau of Economic Research, Cambridge, MA, 2020).

Bloch, C.and Schneider, J. W., "Performance-based funding models and researcher behavior: An analysis of the influence of the norwegian publication indicator at the individual level," Res. Eval. **25**, 371–382 (2016).

Bookstein, A.and Yitzhaki, M., "Own-language preference: A new measure of "relative language self-citation"," Scientometrics **46**, 337–348 (1999).

Bornemann-Cimenti, H., Szilagyi, I. S., and Sandner-Kiesling, A., "Perpetuation of retracted publications using the example of the scott s. reuben case: Incidences, reasons and possible improvements," Sci. Eng. Ethics **22**, 1063–1072 (2016).

Bornmann, L., "Scientific peer review," Annual Rev. Info. Sci & Technol. **45**, 197–245 (2011).

Bornmann, L.and Daniel, H., "What do citation counts measure? a review of studies on citing behavior," Journal of Documentation **64**, 45–80 (2008a).

Bornmann, L.and Daniel, H.-D., "Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by angewandte chemie international edition, or rejected but published elsewhere," J. Am. Soc. Inf. Sci. Technol. **59**, 1841–1852 (2008b).

Bornmann, L.and Leydesdorff, L., "The validation of (advanced) bibliometric indicators through peer assessments: A comparative study using data from InCites and F1000," J. Informetr. **7**, 286–291 (2013).

Brookes, B. C., "The growth, utility, and obsolescence of scientific periodical literature," Journal of Documentation **26**, 283–294 (1970).

Burrel, Q. L., "Stochastic modelling of the first-citation distribution," Scientometrics **52**, 3–12 (2001).

Burrell, Q. L., "The nth-citation distribution and obsolescence," Scientometrics **53**, 309–323 (2002).

Burton, R. E.and Kebler, R. W., "The "half-life" of some scientific and technical literatures," Am. doc. **11**, 18–22 (1960).

Butler, L., "Explaining australia's increased share of ISI publications—the effects of a funding formula based on publication counts," Res. Policy **32**, 143–155 (2003).

Callaham, M., "Journal prestige, publication bias, and other characteristics associated with citation of published studies in Peer-Reviewed journals," JAMA **287**, 2847 (2002).

Caron, E.and Van Eck, N. J., "Large scale author name disambiguation using rule-based scoring and clustering," in *Proceedings of the STI* (2014) pp. 79–86.

Casadevall, A.and Fang, F. C., "Causes for the persistence of impact factor mania," MBio **5**, e00064–14 (2014).

Chatterjee, A., Ghosh, A., and Chakrabarti, B. K., "Universality of citation distributions for academic institutions and journals," PLoS One **11**, e0146762 (2016).

Cole, S., “Professional standing and the reception of scientific discoveries,” Am. J. Sociol. **76**, 286–306 (1970).

Cole, S.and Cole, J. R., “Visibility and the structural bases of awareness of scientific research,” Am. Sociol. Rev. **33**, 397–413 (1968).

Cole, S., Cole, J. R., and Simon, G. A., “Chance and consensus in peer review,” Science **214**, 881–886 (1981).

Dahler-Larsen, P., “Constitutive effects of performance indicators: Getting beyond unintended consequences,” Public Management Review **16**, 969–986 (2014).

De Bellis, N., *Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics* (Scarecrow Press, 2009).

DORA,, “San francisco declaration on research assessment (DORA),” `https://sfdora.org/` (2013), accessed: 2013-NA-NA.

Dorogovtsev, S. N.and Mendes, J. F., “Evolution of networks with aging of sites,” Phys. Rev. E Stat. Phys. Plasmas Fluids Relat. Interdiscip. Topics **62**, 1842–1845 (2000).

Egghe, L., “A theory of continuous rates and applications to the theory of growth and obsolescence rates,” Inf. Process. Manag. **30**, 279–292 (1994).

Egghe, L., *Power Laws in the Information Production Process: Lotkaian Informetrics* (Elsevier Academic Press, Amsterdam, 2005) p. 427.

Egghe, L., “A model showing the increase in time of the average and median reference age and the decrease in time of the price index,” Scientometrics **82**, 243–248 (2010).

Egghe, L.and Ravichandra rao, I. K., “Citation age data and the obsolescence function: Fits and explanations,” Inf. Process. Manag. **28**, 201–217 (1992).

Egghe, L.and Rousseau, R., “Aging, obsolescence, impact, growth, and utilization: Definitions and relations,” (2000).

Eisen, J. A., MacCallum, C. J., Neylon, C., Sugimoto, C. R., and Walport, M., “Expert failure: Re-evaluating research assessment,” PLoS Biol. **11**, e1001677 (2013).

Ellis, L. V.and Durden, G. C., “Why economists rank their journals the way they do,” J. Econ. Bus. **43**, 265–270 (1991).

Eom, Y.-H.and Fortunato, S., “Characterizing and modeling citation dynamics,” PLoS One **6**, e24926 (2011).

Ernst, E., Saradeth, T., and Resch, K. L., “Drawbacks of peer review,” Nature **363**, 296 (1993).

Evans, J. a., “Electronic publication and the narrowing of science and scholarship,” Science **321**, 395–399 (2008).

Evans, J. A.and Foster, J. G., “Metaknowledge,” Science **331**, 721–725 (2011).

Eyre-Walker, A.and Stoletzki, N., “The assessment of science: The relative merits of Post-Publication review, the impact factor, and the number of citations,” PLoS Biol. **11**, e1001675 (2013).

Forscher, P. S., Brauer, M., Azevedo, F., Cox, W. T. L., and Devine, P. G., “How many reviewers are required to obtain reliable evaluations of NIH R01 grant proposals?” (2019).

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., and Barabási, A.-L., “Science of science,” Science **359** (2018), 10.1126/science.aao0185.

Furman, J. L., Jensen, K., and Murray, F., “Governing knowledge in the scientific community: Exploring the role of retractions in biomedicine,” Res. Policy **41**, 276–290 (2012).

Garfield, E., “The obliteration phenomenon in science, and the advantage of being obliterated,” Current Contents , 5–7 (1957).

Gingras, Y., Larivière, V., Macaluso, B., and Robitaille, J.-P., “The effects of aging on researchers’ publication and citation patterns,” PLoS One **3**, e4048 (2008).

Gläser, J., “The social orders of research evaluation systems,” in *The Changing Governance of the Sciences: The Advent of Research Evaluation Systems*, edited by R. Whitley and J. Gläser (Springer Netherlands, Dordrecht, 2007) pp. 245–266.

Gläser, J.and Laudel, G., “Governing science,” Arch. Eur. Sociol. **57**, 117–168 (2016).

Goldberg, S. R., Anthony, H., and Evans, T. S., “Modelling citation networks,” Scientometrics **105**, 1577–1604 (2015).

Goodman, S. N., Berlin, J., Fletcher, S. W., and Fletcher, R. H., “Manuscript quality before and after peer review and editing at annals of internal medicine,” Ann. Intern. Med. **121**, 11 (1994).

Gross, P. L.and Gross, E. M., “COLLEGE LIBRARIES AND CHEMICAL EDUCATION,” Science **66**, 385–389 (1927).

Hajra, K. B.and Sen, P., “Modelling aging characteristics in citation networks,” Physica A: Statistical Mechanics and its Applications **368**, 575–582 (2006).

Hargens, L. L.and Felmlee, D. H., “Structural determinants of stratification in science,” Am. Sociol. Rev. **49**, 685 (1984).

Herman, E., “Research in progress. part 2 – some preliminary insights into the information needs of the contemporary academic researcher,” Aslib Proc. **56**, 118–131 (2004a).

Herman, E., “Research in progress: some preliminary and key insights into the information needs of the contemporary academic researcher. part 1,” Aslib Proc. **56**, 34–47 (2004b).

Hicks, D., “Performance-based university research funding systems,” Res. Policy **41**, 251–261 (2012).

Hicks, D., Wouters, P., Waltman, L., Rijcke, S. d., and Rafols, I., “The leiden manifesto for research metrics,” Nature **520**, 429–431 (2015).

Higham, K. W., Governale, M., Jaffe, A. B., and Zülicke, U., “Unraveling the dynamics of growth, aging and inflation for citations to scientific articles from specific research fields,” Journal of Informetrics **11**, 1190–1200 (2017), arXiv:1708.08335 [cs.DL].

Hood, W. W.and Wilson, C. S., “The literature of bibliometrics, scientometrics, and informetrics,” Scientometrics **52**, 291–314 (2001).

King, S. R. F., “Consultative review is worth the wait,” Elife **6**, e32012 (2017).

Klebel, T.and Traag, V., “Introduction to causality in science studies,” SocArXiv (2024), 10.31235/osf.io/4bw9e.

Klein, M., Broadwell, P., Farb, S. E., and Grappone, T., “Comparing published scientific journal articles to their pre-print versions,” in *Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries - JCDL ’16* (ACM Press, New York, New York, USA, 2016) pp. 153–162.

Knothe, G., “Comparative citation analysis of duplicate or highly related publications,” J. Am. Soc. Inf. Sci. Technol. **57**, 1830–1839 (2006).

Kuhn, T. S., *The structure of scientific revolutions* (The University of Chicago Press, Chicago, 2012).

Laherrère, J.and Sornette, D., “Stretched exponential distributions in nature and economy: “fat tails” with characteristic scales,” The European Physical Journal B - Condensed Matter and Complex Systems **2**, 525–539 (1998).

Lai, K. H., Traag, V., and Waltman, L., “Challenges in using bibliometric indicators to assess peer review decisions: A simulation model,” in *PEERE Conference* (2020).

Lamers, W. S., Boyack, K., Larivière, V., Sugimoto, C. R., and others,, “Meta-Research: Investigating disagreement in the scientific literature,” Elife (2021).

Larivière, V., Archambault, É., and Gingras, Y., “Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004),” J. Am. Soc. Inf. Sci. Technol. **59**, 288–296 (2008).

Larivière, V.and Gingras, Y., “The impact factor’s matthew effect: A natural experiment in bibliometrics,” J. Am. Soc. Inf. Sci. Technol. **61**, 424–427 (2010).

Larivière, V., Gingras, Y., and Archambault, É., “The decline in the concentration of citations, 1900-2007,” J. Am. Soc. Inf. Sci. Technol. **60**, 858–862 (2009).

Larivière, V., Gingras, Y., Sugimoto, C. R., and Tsou, A., “Team size matters: Collaboration and scientific impact since 1900: On the relationship between collaboration and scientific impact since 1900,” J. Assoc. Inf. Sci. Technol. **66**, 1323–1332 (2015).

Lariviere, V., Kiermer, V., MacCallum, C. J., McNutt, M., Patterson, M., Pulverer, B., Swaminathan, S., Taylor, S., and Curry, S., "A simple proposal for the publication of journal citation distributions," bioRxiv , 062109 (2016).

Larivière, V.and Sugimoto, C. R., "The journal impact factor: A brief history, critique, and discussion of adverse effects," in *Springer Handbook of Science and Technology Indicators*, edited by W. G. F. M. S. Thelwall (Springer, Cham, 2019) pp. 3–24.

Lee, C. J., Sugimoto, C. R., Zhang, G., and Cronin, B., "Bias in peer review," J. Am. Soc. Inf. Sci. Technol. **64**, 2–17 (2013).

Line, M. B., "THE 'HALF-LIFE' OF PERIODICAL LITERATURE: APPARENT AND REAL OBSOLESCENCE," Journal of Documentation **26**, 46–54 (1970).

Line, M. B.and Sandison, A., "PROGRESS IN DOCUMENTATION: 'obsolescence' and changes in the use of literature with time," Journal of Documentation **30**, 283–350 (1974).

Lozano, G. A., Larivière, V., and Gingras, Y., "The weakening relationship between the impact factor and papers' citations in the digital age," J. Am. Soc. Inf. Sci. Technol. (2012), 10.1002/asi.22731.

MacRoberts, M. H.and MacRoberts, B. R., "Problems of citation analysis: A critical review," Journal of the American Society for Information Science (1986-1998) **40**, 342 (1989).

McCain, K. W., "Eponymy and obliteration by incorporation: The case of the "nash equilibrium"," J. Am. Soc. Inf. Sci. Technol. **62**, 1412–1424 (2011).

McKiernan, E. C., Schimanski, L. A., Muñoz Nieves, C., Matthias, L., Niles, M. T., and Alperin, J. P., "Use of the journal impact factor in academic review, promotion, and tenure evaluations," Elife **8** (2019), 10.7554/eLife.47338.

Medoff, M. H., "Evidence of a harvard and chicago matthew effect," Journal of Economic Methodology **13**, 485–506 (2006).

Merton, R. K., "The matthew effect in science," Science **159**, 56–63 (1968).

Milojević, S., Radicchi, F., and Bar-Ilan, J., "Citation success index - an intuitive pair-wise journal comparison metric," Journal of Informetrics **11**, 223–231 (2016), arXiv:1607.03179 [cs.DL].

Mingers, J.and Burrell, Q. L., "Modeling citation behavior in management science journals," Inf. Process. Manag. **42**, 1451–1464 (2006).

Mingers, J.and Xu, F., "The drivers of citations in management science journals," Eur. J. Oper. Res. **205**, 422–430 (2010).

Moed, H. F., "UK research assessment exercises: Informed judgments on research quality or quantity?" Scientometrics **74**, 153–161 (2008).

Molas-Gallart, J.and Rafols, I., "Why bibliometric indicators break down: Unstable parameters, incorrect models and irrelevant properties," (2018).

Moreira, J. A. G., Zeng, X. H. T., and Amaral, L. A. N., "The distribution of the asymptotic number of citations to sets of publications by a researcher or from an academic department are consistent with a discrete lognormal model," PLoS One **10**, e0143108 (2015).

Müller, R.and de Rijcke, S., "Thinking with indicators. exploring the epistemic impacts of academic performance indicators in the life sciences," Res. Eval. **26**, 157–168 (2017).

Nakamoto, H., "Synchronous and diachronous citation distributions," Informetrics (1988).

Okike, K., Hug, K. T., Kocher, M. S., and Leopold, S. S., "Single-blind vs double-blind peer review in the setting of author prestige," JAMA **316**, 1315–1316 (2016).

Olensky, M., Schmidt, M., and Van Eck, N. J., "Evaluation of the citation matching algorithms of CWTS and iFQ in comparison to the web of science," Journal of the Association for Information Science and Technology **67**, 2550–2564 (2016).

O'Neil, C., *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown, 2016).

Osuna, C., Cruz-Castro, L., and Sanz-Menéndez, L., "Overturning some assumptions about the effects of evaluation systems on publication performance," Scientometrics **86**, 575–592 (2011).

Owens, B., "Research assessments: Judgement day," Nature **502**, 288–290 (2013).

Pan, R. K., Kaski, K., and Fortunato, S., "World citation and collaboration networks: uncovering the role of geography in science," Scientific Reports 2012 2 **2**, 902 (2012).

Pan, R. K., Petersen, A. M., Pammolli, F., and Fortunato, S., "The memory of science: Inflation, myopia, and the knowledge network," J. Informetr. **12**, 656–678 (2018).

Parolo, P. D. B., Pan, R. K., Ghosh, R., Huberman, B. A., Kaski, K., and Fortunato, S., "Attention decay in science," J. Informetr. **9**, 734–745 (2015).

Penner, O., Pan, R. K., Petersen, A. M., Kaski, K., and Fortunato, S., "On the predictability of future impact in science," Sci. Rep. **3** (2013), 10.1038/srep03052.

Peritz, B. C., "On the association between journal circulation and impact factor," J. Inf. Sci. Eng. **21**, 63–67 (1995).

Perneger, T. V., "Citation analysis of identical consensus statements revealed journal-related bias," J. Clin. Epidemiol. **63**, 660–664 (2010).

Peterson, G. J., Pressé, S., and Dill, K. A., "Nonuniversal power law scaling in the probability distribution of scientific citations," Proc. Natl. Acad. Sci. U. S. A. **107**, 16023–16027 (2010).

Phillips, D. P., Kanter, E. J., Bednarczyk, B., and Tastad, P. L., "Importance of the lay press in the transmission of medical knowledge to the scientific community," N. Engl. J. Med. **325**, 1180–1183 (1991).

Pier, E. L., Brauer, M., Filut, A., Kaatz, A., Raclaw, J., Nathan, M. J., Ford, C. E., and Carnes, M., "Low agreement among reviewers evaluating the same NIH grant applications," Proc. Natl. Acad. Sci. U. S. A. **115**, 2952–2957 (2018).

Poncela-Casasnovas, J., Gerlach, M., Aguirre, N., and Amaral, L. A. N., "Large-scale analysis of micro-level citation patterns reveals nuanced selection criteria," Nature Human Behaviour **3**, 568–575 (2019).

Price, D. D. S., "A general theory of bibliometric and other cumulative advantage processes," Journal of the American Society for Information Science **27**, 292–306 (1976).

Price, D. J., "NETWORKS OF SCIENTIFIC PAPERS," Science **149**, 510–515 (1965).

Radicchi, F.and Castellano, C., "Rescaling citations of publications in physics," Physical Review E **83**, 046116 (2011).

Radicchi, F.and Castellano, C., "A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions," PLoS One **7**, e33833 (2012).

Radicchi, F., Fortunato, S., and Castellano, C., "Universality of citation distributions: toward an objective measure of scientific impact," Proc. Natl. Acad. Sci. U. S. A. **105**, 17268–17272 (2008).

Radicchi, F., Weissman, A., and Bollen, J., "Quantifying perceived impact of scientific publications," J. Informetr. **11**, 704–712 (2017).

Redner, S., "How popular is your paper? an empirical study of the citation distribution," Eur. Phys. J. B **4**, 131–134 (1998).

Redner, S., "Citation statistics from 110 years of physical review," Phys. Today **58**, 49–54 (2005).

de Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., and Hammarfelt, B., "Evaluation practices and effects of indicator use—a literature review," Res. Eval. **25**, 161–169 (2016).

Van de Rijt, A., Kang, S. M., Restivo, M., and Patil, A., "Field experiments of success-breeds-success dynamics," Proc. Natl. Acad. Sci. U. S. A. **111**, 6934–6939 (2014).

Rothwell, P. M.and Martyn, C. N., "Reproducibility of peer review in clinical neuroscience. is agreement between reviewers any greater than would be expected by chance alone?" Brain **123 (Pt 9)**, 1964–1969 (2000).

Rushforth, A.and de Rijcke, S., "Accounting for impact? the journal impact factor and the making of biomedical research in the netherlands," Minerva **53**, 117–139 (2015).

Sandström, U.and Van den Besselaar, P., "Funding, evaluation, and the performance of national research systems," J. Informetr. **12**, 365–384 (2018).

Schneider, J. W., Aagaard, K., and Bloch, C. W., "What happens when national research funding is linked to differentiated publication counts? a comparison of the australian and norwegian publication-based funding models," Res. Eval. **25**, 244–256 (2016).

Schubert, A.and Glänzel, W., "Cross-national preference in coauthorship, references and citations," Scientometrics **69**, 409–428 (2006).

Seeber, M., Cattaneo, M., Meoli, M., and Malighetti, P., "Self-citations as strategic response to the use of metrics for career decisions," Res. Policy (2017), 10.1016/j.respol.2017.12.004.

Seglen, P. O., "Causal relationship between article citedness and journal impact," Journal of the American Society for Information Science **45**, 1–11 (1994).

Seglen, P. O., "Why the impact factor of journals should not be used for evaluating research," BMJ **314**, 498–502 (1997).

Simkin, M. V.and Roychowdhury, V. P., "A mathematical theory of citing," J. Am. Soc. Inf. Sci. Technol. **58**, 1661–1673 (2007).

Sinatra, R., Deville, P., Szell, M., Wang, D., and Barabási, A. L., "A century of physics," Nat. Phys. **11**, 791–796 (2015).

Sinatra, R., Wang, D., Deville, P., Song, C., and Barabási, A.-L., "Quantifying the evolution of individual scientific impact," Science **354** (2016), 10.1126/science.aaf5239.

Smaldino, P. E.and McElreath, R., "The natural selection of bad science," Royal Society Open Science **3**, 160384 (2016).

Starbuck, W. H., "How much better are the Most-Prestigious journals? the statistics of academic publication," Organization Science **16**, 180–200 (2005).

Stegehuis, C., Litvak, N., and Waltman, L., "Predicting the long-term citation impact of recent publications," J. Informetr. **9**, 642–657 (2015).

Stinson, E. R.and Lancaster, F. W., "Synchronous versus diachronous methods in the measurement of obsolescence by citation studies," J. Inf. Sci. Eng. **13**, 65–74 (1987).

Stringer, M. J., Sales-Pardo, M., and Amaral, L. A. N., "Effectiveness of journal ranking schemes as a tool for locating information," PLoS One **3**, e1683 (2005).

Šubelj, L.and Fiala, D., "Publication boost in web of science journals and its effect on citation distributions," J. Assoc. Inf. Sci. Technol. **68**, 1018–1023 (2017).

Sugimoto, C. R.and Larivière, V., *Measuring Research: What Everyone Needs to Know* (Oxford University Press, 2018) pp. 1–143.

Teplitskiy, M., Duede, E., Menietti, M., and Lakhani, K. R., "Status drives how we cite: Evidence from thousands of authors," , 20 (2020), arXiv:2002.10033 [cs.SI].

Thelwall, M.and Wilson, P., "Regression for citation data: An evaluation of different methods," J. Informetr. **8**, 963–971 (2014).

Tiokhin, L., Panchanathan, K., Smaldino, P. E., and Lakens, D., "Shifting the level of selection in science," (2021).

Tomkins, A., Zhang, M., and Heavlin, W. D., "Reviewer bias in single- versus double-blind peer review," Proc. Natl. Acad. Sci. U. S. A. **114**, 12708–12713 (2017).

Traag, V. A., "Inferring the causal effect of journals on citations," Quantitative Science Studies , 1–9 (2021).

Traag, V. A., Malgarini, M., and Sarlo, S., "Metrics and peer review agreement at the institutional level," (2020), arXiv:2006.14830 [cs.DL].

Traag, V. A.and Waltman, L., "Systematic analysis of agreement between metrics and peer review in the UK REF," Palgrave Communications **5**, 29 (2019).

Traag, V. A.and Waltman, L., "Causal foundations of bias, disparity and fairness," arXiv (2022), 10.48550/arXiv.2207.13665.

Tsallis, C.and de Albuquerque, M. P., "Are citations of scientific papers a case of nonextensivity?" Eur. Phys. J. B **13**, 777–780 (2000).

Verstak, A., Acharya, A., Suzuki, H., Henderson, S., Iakhiaev, M., Lin, C. C. Y., and Shetty, N., "On the shoulders of giants: The growing impact of older articles," (2014), arXiv:1411.0275 [cs.DL].

Vinkler, P., "Relationships between the rate of scientific development and citations. the chance for citedness model," Scientometrics **35**, 375–386 (1996).

Visser, M., Van Eck, N. J., and Waltman, L., "Large-scale comparison of bibliographic data sources: Scopus, web of science, dimensions, crossref, and microsoft academic," Quantitative Science Studies **2**, 20–41 (2021).

Wallace, M. L., Larivière, V., and Gingras, Y., "Modeling a century of citation distributions," J. Informetr. **3**, 296–303 (2009).

Waltman, L.and Traag, V. A., "Use of the journal impact factor for assessing individual articles: Statistically flawed or not?" F1000Res. **9**, 366 (2020), arXiv:1703.02334 [cs.DL].

Waltman, L.and Van Eck, N. J., "Field normalization of scientometric indicators," Springer Handbook of Science and Technology Indicators , 281–300 (2019).

Waltman, L., Van Eck, N. J., and Van Raan, A. F. J., "Universality of citation distributions revisited," J. Am. Soc. Inf. Sci. Technol. **63**, 72–77 (2011).

Wang, D.and Barabási, A.-L., *The Science of Science*, 1st ed. (Cambridge University Press, 2021).

Wang, D., Song, C., and Barabási, A.-L., "Quantifying Long-Term scientific impact," Science **342**, 127–132 (2013).

Wang, D., Song, C., Shen, H.-W., and Barabási, A.-L., "Science communication. response to comment on "quantifying long-term scientific impact"," Science **345**, 149–149 (2014).

Wang, J., Mei, Y., and Hicks, D., "Comment on "quantifying long-term scientific impact"," Science **345**, 149–149 (2014).

Wang, M., Yu, G., and Yu, D., "Effect of the age of papers on the preferential attachment in citation networks," Physica A: Statistical Mechanics and its Applications **388**, 4273–4276 (2009).

Way, S. F., Morgan, A. C., Larremore, D. B., and Clauset, A., "Productivity, prominence, and the effects of academic environment," Proc. Natl. Acad. Sci. U. S. A. **166**, 10729–10733 (2019).

Wilsdon, J., "We need a measured approach to metrics," Nature **523**, 129 (2015).

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., and Johnson, B., "Metric tide: Report of the independent review of the role of metrics in research assessment and management," Tech. Rep. (Higher Education Funding Council for England, 2015).

Woods, H. B., Brumberg, J., Kaltenbrunner, W., Pinfield, S., and Waltman, L., "Innovations in peer review in scholarly publishing: a meta-summary," (2022).

Wouters, P., Sugimoto, C. R., Larivière, V., McVeigh, M. E., Pulverer, B., de Rijcke, S., and Waltman, L., "Rethinking impact factors: better ways to judge a journal," Nature **569**, 621–623 (2019).

Wu, L., Wang, D., and Evans, J. A., "Large teams develop and small teams disrupt science and technology," Nature **566**, 378–382 (2019).

Yin, Y.and Wang, D., "The time dimension of science: Connecting the past to the future," Journal of Informetrics **11**, 608–621 (2017), arXiv:1704.04657 [physics.soc-ph].

Zeng, A., Shen, Z., Zhou, J., Wu, J., Fan, Y., Wang, Y., and Stanley, H. E., "The science of science: From the perspective of complex systems," Phys. Rep. **714-715**, 1–73 (2017).

Zhang, L., Rousseau, R., and Sivertsen, G., "Science deserves to be judged by its contents, not by its wrapping: Revisiting seglen's work on journal impact and research evaluation," PLoS One **12**, e0174205 (2017).