Statistical Validation of a Global Model for the Distribution of the Ultimate Number of Citations Accrued by Papers Published in a Scientific Journal

Michael J. Stringer
Department of Physics and Astronomy, Northwestern University, Evanston, IL 60208
Northwestern Institute on Complex Systems (NICO), Northwestern University, Evanston, IL 60208.
E-mail: m-stringer@northwestern.edu

Marta Sales-Pardo
Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208
Northwestern Institute on Complex Systems (NICO), Northwestern University, Evanston, IL 60208
Departament d'Enginyeria Química, Universitat Rovira i Virgili, Tarragona, 43007, Spain.

Luís A. Nunes Amaral
Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208
Howard Hughes Medical Institute, Northwestern University, Evanston, IL 60208.

A central issue in evaluative bibliometrics is the characterization of the citation distribution of papers in the scientific literature. Here, we perform a large-scale empirical analysis of journals from every field in Thomson Reuters' Web of Science database. We find that only 30 of the 2,184 journals have citation distributions that are inconsistent with a discrete lognormal distribution at the rejection threshold that controls the false discovery rate at 0.05. We find that large, multidisciplinary journals are over-represented in this set of 30 journals, leading us to conclude that, within a discipline, citation distributions are lognormal. Our results strongly suggest that the discrete lognormal distribution is a globally accurate model for the distribution of "eventual impact" of scientific papers published in single-discipline journal in a single year that is removed sufficiently from the present date.

Introduction

Citation analysis is a widely-used approach for filtering scientific information. The growth of the R&D workforce, the number of scientific fields, and the number and size of data repositories of research output (Tomlin, 2005) suggest that citation analysis, and other automatic methods of research classification and assessment, are likely to become even more widespread. Despite its increasing usage and importance, there remains deep distrust of citation analysis within the broad scientific community ("Experts still needed", 2009, "Not-So-Deep Impact", 2005, Adam, 2002). At the extreme, some critics claim that "the practice is so riddled with errors and biases that it can be worse than useless" (Adam, 2002, p. 729). Thus, it is important to develop methods of citation analysis that reduce errors and biases and that are informed by empirical patterns of citations to scholarly publications (Lane, 2009).

The methodological criticisms of citation-based research evaluation typically concern the following three broad aspects of empirically observed patterns of citation to papers:

1. Scientific fields have heterogeneous citation properties, so comparison across fields is unwarranted. For example, computer scientists likely have different publication practices and adhere to different citation norms than sociologists (Wouters, 1999). Although there is much research devoted to defining exactly what a field is, as well as identifying fields using citation data (Shiffrin & Borner, 2004, Boyack, Klavans, & Börner, 2005), the definition and identification of fields remain a difficult problem. An analysis that aims to compare papers, even implicitly in the form of studying a citation distribution, must include only comparable papers to be interpretable in a straightforward manner.
2. Citation counts are dynamic, so evaluations are made before all the information is in. Indeed, the citation count of many papers is still increasing, and almost never assured to remain fixed (Glänzel & Garfield, 2004, Burrell,

Received December 17, 2009; revised February 11, 2010; accepted February 11, 2010

 DOI: 10.1002/asi.21335

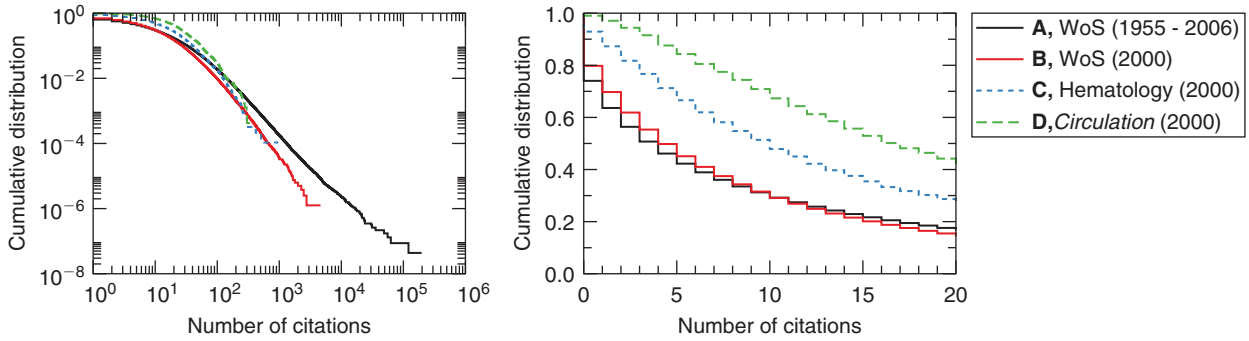

FIG. 1. Citation distributions differ by level of aggregation, as well as time period. Citations for all papers in these sets are tabulated at the end of 2006. The left panel is on double logarithmic scales, and the right panel is the same data on linear scale. A, Distribution of the number of citations to all scientific articles indexed in the Web of Science between 1955 and 2006. B, Distribution of citations to all scientific articles published in calendar year 2000. Note that the tail decays faster, for high impact papers have not yet had enough time to accumulate citations. C, Distribution of citations to all scientific articles published in year 2000 in the field of "Hematology." Note that the median number of citations is significantly ($p < 0.001$) higher in hematology than the median number of citations overall. D, Distribution of citations to all scientific articles published in year 2000 in the journal *Circulation* (which is classified in the field of "Hematology"). The median number of citations to papers published in *Circulation* is significantly higher ($p < 0.001$) than the median number of citations to papers in hematology. At the aggregation level of D, we find that almost all of the data is consistent with a discrete lognormal distribution. Thus, the global distribution A is likely a mixture of discrete lognormal distributions.

2003, Egghe & Rousseau, 2000). Counting the number of citations after a fixed time period biases metrics in favor of fields where paper typically accumulate citations more rapidly. Also, the functional form of the citation distribution depends on the set of years included in the analysis (Simkin & Roychowdhury, 2007). For example, under a simple cumulative advantage model for citation network growth, the global distribution of citations to papers is a power law, but the distribution of citations to papers published within the same year is exponential (Simkin & Roychowdhury; Krapivsky & Redner, 2001).

3. Citation distributions are skewed, so averages are heavily influenced by extreme values. The skewness of bibliometric distributions, especially the skewness in the distribution of the number of citations to papers, has been extensively studied as far back as the 1920's (Lotka, 1926). However, empirical reports of the distribution of the number of citations are often conflicting and depend on the level of aggregation of the analysis (see Figure 1). The models typically used are the power-law distribution (Lotka; Solla Price, 1976; Redner, 1998) and the lognormal distribution (Stewart, 1994; Redner, 2005; Stringer, Sales-Pardo, & Amaral, 2008; Radicchi, Fortunato, & Castellano, 2008), but a number of other skewed distributions have also been considered (Nadarajah & Kotz, 2007). For any of these distributions, the average of a sample is a misleading indicator of the typical value.

Stringer et al. (2008) accounted for these methodological challenges by (a) restricting analysis to papers published in the same journal and year, (b) focusing on the logarithm of the number of citations, and (c) observing the distribution of the number of citations to journal after the papers have stopped being cited. In order to make sure that papers in a set are comparable, Stringer et al. considered separately papers published in the same journal and year, because as one of the primary reasons that journals exists is to group papers by topical relevance (Cole, 2000). An exception to this rule may be large, multidisciplinary journals; they may publish papers that are of interest to a broad readership as opposed to papers that are explicitly related by subject matter. Nevertheless, for a large majority of journals, we expect that papers published within a journal will be comparable in citation properties.

Stringer et al. (2008) showed that within most journals, the distribution of the number of citations to papers within the journal has a characteristic time period, after which the distribution is no longer changing, i.e., the papers in a journal are no longer being cited to an appreciable extent. This suggests that one way to eliminate the confounding issue of citation dynamics is to consider how many citations a paper has accumulated after it has stopped being read and cited. Indeed, the ultimate number of citations that a paper receives may be a more intuitive interpretation of the impact that a paper had on the research community.

Stringer et al. (2008) also showed that a lognormal provides good visual agreement with the citation distribution within journals. Indeed, as shown in Figure 2 of Stringer et al. (2008), the large skewness and kurtosis of the data for all journals lead to the immediate rejection of the hypothesis that the data can be described by classical distributions such as the Gaussian, exponential, or double-exponential.

Moreover, there are good a priori reasons to investigate the lognormal distribution as a candidate for the distribution of citations to papers in scientific journals. A lognormal model for the distribution for the number of citations to papers is a plausible model if one assumes that the number of citations that a paper receives depends exponentially on a hidden quantity that aggregates several factors, and that a weakness in any one factor reduces the effect of all the other factors (Stewart, 1994). For example, if a paper must be relevant to current

DOI: 10.1002/asi

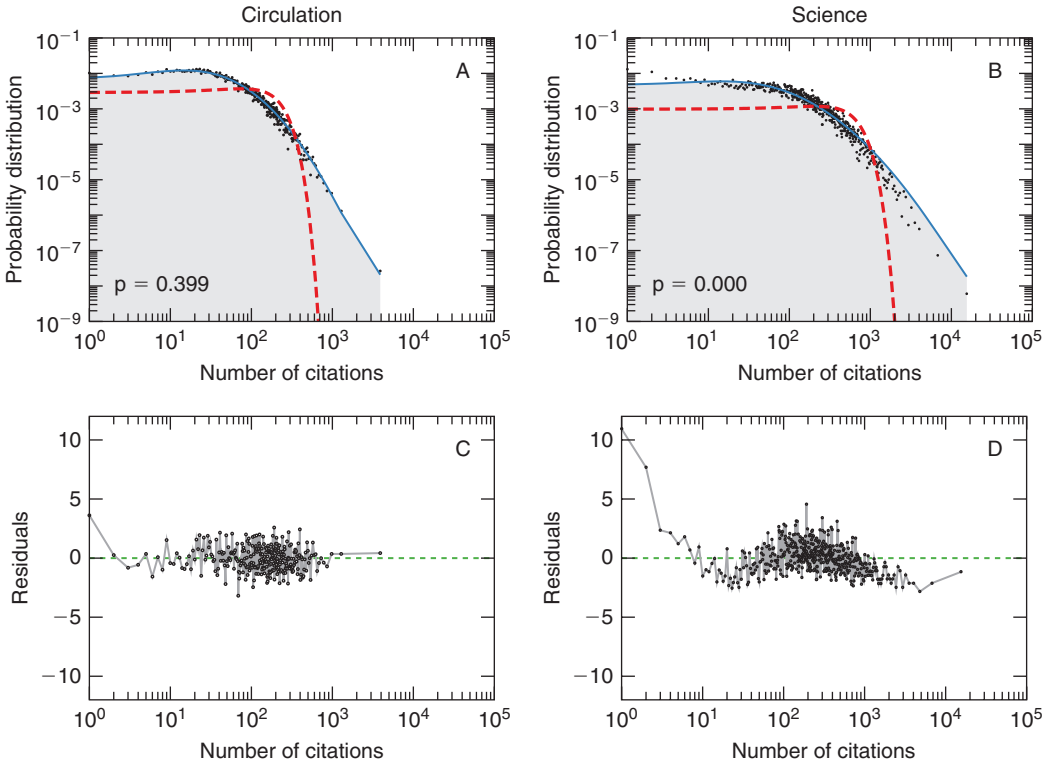

FIG. 2. Goodness of fit of latent normal model to journal citation distribution data.
A, Comparison of the model to data for articles published in the steady state period (1970–1998) for the journal *Circulation*. We can not reject hypothesis H_1 ($p_1 = 0.4$). B, Plot of residuals, $\chi = \frac{E-A}{\sqrt{E}}$ against the independent variable, n for the journal *Circulation*. For journals where hypothesis bfH_1 cannot be rejected ($p_1^j > 0.05$), the residuals are uncorrelated with the number of citations. C, Comparison of the model to data for articles published in the steady state period (1996–1998) of the journal *Science*. In this case, p_1 is near zero, indicating that we can reject hypothesis H_1 with high confidence. D, Plot of residuals, $\chi = \frac{E-A}{\sqrt{E}}$ against the independent variable, n for the journal *Science*. For journal where hypothesis H_1 is rejected ($p_1^j < 0.05$), the residuals are correlated. In this particular case, the model under-predicts the number of uncited articles. Whereas for the "true" model of all journal citation distributions, we would expect 5% (109) of the journals to yield $p_1^j > 0.05$, we observe that 10% (229) of the journals yield $p_1^j > 0.05$. For the purpose of comparison, the dashed lines in A and C indicate best fit normal distributions.

research *and* technically sound *and* visible to the research community *and* clearly written, and so on, then one might expect the hidden variable to be normally distributed and the number of citations to be lognormally distributed (Shockley, 1957).

Here, we use statistical hypothesis testing methods to show that the empirically observed citation distribution is consistent with a discrete lognormal distribution for all but 30 of the 2,184 journals in the Thomson Reuters' Web of Science (WoS) database at an appropriate rejection threshold for multiple tests. For each of the 30 journals that are inconsistent with the discrete lognormal model, we investigate the reasons for the inconsistency. Our findings indicate that for 23 of the 30 journals that fail, the inconsistency is because of the fact that the journal citation distribution is changing "enough" over time to be detected by the statistical test. The seven remaining journals are primarily large, multidisciplinary journals and, thus, likely contain a mixture of papers from different fields with different citation properties. Our results strongly support the hypothesis that a globally accurate model for the distribution of "eventual impact" of scientific papers published in the same journal and year is a discrete lognormal distribution.

Methods

Data

We studied citation data from papers in the WoS database, which we gathered using a Web interface available to those with a subscription to the service from Thomson Reuters (http://www.isiwebofknowledge.com). Within the WoS database, we examined papers in the Science Citation Index published during 1955–2006, the Social Science Citation Index published during 1956–2006, and the Arts & Humanities Citation Index published during 1975–2006. All citations counts were enumerated as of December 31, 2006.

To ensure that we do not mix results from different types of published literature, we restrict the analysis to primary literature, which we identify using a series of filters to restrict the papers that we analyze. Before applying any filters, there were 36,658,661 publications assigned to 16,320journals.[1]

[1] "Journal" is taken as those papers that have the same journal abbreviation in the WoS database. This is the practice that is followed in Journal Citation Reports, and helps aggregate the data correctly.

DOI: 10.1002/asi

We first restrict the papers that we analyze to those classified with a document type of "Article" in the WoS database, which reduces the number of publications to 22,951,535 articles and 16,117 journals. We then further restrict our attention to papers published in journals that contain at least 50 articles per year during, at least, 15 years. This ensures that we can implement the procedure described in Stringer et al. (2008) to identify steady-state periods for the distribution of the number of citations to papers in that journal and analyze the aggregate distribution of citations to papers published during that period. Finally, we consider only journals in which fewer than 75% of the papers remain uncited in the long run—the 68 journals in which more than 75% of papers remain uncited are not primary research literature; they are trade journals such as *Dr. Dobbs Journal*, science news magazines such as *The Scientist*, or non-English language journals that have poor coverage in the WoS database, such as *Measurement Techniques-USSR Journal*.

After filtering, we are left with 12,454,829 primary research articles assigned to 2,184 journals. There is at least one journal included in our analysis from 213 of the 220 fields represented in the 2006 version of the *Journal Citation Reports*. The fields that are not represented are relatively new and do not include any journals that have reached a steady-state distribution. Most of the papers excluded by the filtering process are neither primary literature nor in journals that have been indexed by the WoS long enough to reach a steady-state citation distribution. The remaining excluded publications are in low-volume journals and journals that no longer exist. We find no reason to believe that the low-volume or newly created journals that we exclude from our analysis would exhibit different behavior from the journals that we study.

Identification of the Steady-State Distribution

The distribution of the number of citations that papers have received changes in time, as papers accumulate citations. To eliminate this confounding effect, we use the heuristic method described in Stringer et al. (2008) to identify periods of time for each journal where the yearly citation distributions are statistically identical, that is, papers in that journal are no longer getting cited enough to change the citation distribution significantly. In cases where we identify multiple steady-state periods in the history of a journal, such as *Ecology*, see Figure 4 of Stringer et al. (2008), we consider only the most recent steady-state period.

Discrete Lognormal Model

One problem with using a lognormal distribution to model citations to papers is that a lognormal distribution is defined over positive real numbers, whereas citation counts are non-negative integers. There have been a number of ways that researchers have modified the lognormal distribution to account for the fact that citation counts are discrete and include zero counts. Often zero counts are excluded, but this is not appropriate because receiving zero citations is the single most common outcome. Another modification is the delta-lognormal distribution (Aitchison, 1955), which modifies the standard lognormal distribution by allowing a fraction of the papers to have zero citations. In a different approach, Stewart (1994) assumes that citation counts are a result of a Poisson process, with citation rates that are lognormally distributed.

Following Stringer et al. (2008), we make the conversion from a continuous lognormal distribution to a discrete version in the following way. We surmise that the number of citations a paper receives is the result of a latent variable, q, and, thus, any paper with a value of q in some range will receive n citations (Burrell, 2001). In this model, the continuous lognormal distribution would be written as $n_{LN}(q) = 10^q$, where we assume that q is normally distributed, $q \sim N(\mu, \sigma)$. Perhaps the "simplest" way of mapping the continuous value of q onto the discrete value of n is just rounding to the nearest integer. However, because q is unobservable, it is not clear what value of n should result of a given q. Thus, we introduce a parameter g that allows for a discrete mapping that includes the simple floor function ($\gamma = 0$) and rounding to the nearest integer ($\gamma = -\frac{1}{2}$). The parameter γ can be interpreted as the value of q needed for a paper to get one citation. This assumption leads to the following form for the citation distribution:

$$p(n|\mu,\sigma,\gamma) = \begin{cases} \int_{-\infty}^{\log_{10}(\gamma+1)} \frac{dq}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(q-\mu)^2}{2\sigma^2}\right) & n = 0 \\ \int_{\log_{10}(n+\gamma)}^{\log_{10}(n+\gamma+1)} \frac{dq}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(q-\mu)^2}{2\sigma^2}\right) & n \geq 1 \end{cases} \quad (1)$$

We note that for large values of q, the change as a result of discretization is negligible, and the distribution is nearly identical to the lognormal distribution evaluated at only integer values. However, when the mean of the underlying normal distribution is small, the distribution is significantly different from a continuous lognormal distribution. For example, if $\mu < 0$, the distribution $p(n)$ has a mode at zero and is monotonically decreasing.

Parameter Estimation Procedure

In the past, one obstacle to using a model where the probability distribution cannot be written in closed form is the prohibitive computational effort required to estimate the parameters of the model. Fortunately, it is now computationally feasible to estimate parameters using maximum likelihood estimation (MLE) in a straightforward manner. MLE is justified in this case because we do not have any prior knowledge about what values the parameters are likely to take for a given journal. To proceed, we, thus, assume that within a journal, n is identically and independently distributed. Because the logarithm is a monotonically increasing function, then maximizing the log-likelihood, $\mathcal{L} = \log p(\{n\}|\mu,\sigma,\gamma)$, assures we are obtaining the most likely values of the parameters given the observed data.

DOI: 10.1002/asi

The application of the MLE formalism to this particular case yields:

$$\mathcal{L} = f(0)\log \int_{-\infty}^{\log(\gamma+1)} dq\ N(q|\mu,\sigma) + \sum_{n=1}^{\infty} f(n)\log \int_{\log(n+\gamma)}^{\log(n+\gamma+1)} dq\ N(q|\mu,\sigma), \quad (2)$$

where $f(n)$ is the number of papers receiving n citations. For each journal, we numerically find the parameter values that maximize $\mathcal{L}$ using the downhill simplex algorithm (Press, Teukolsky, Vetterling, & Flannery, 2002).

Hypothesis Testing Procedures

We test several thousand hypotheses in our study. For such multiple testing situations, one must be careful about the rejection threshold used, as well as the statistical testing methodology (Benjamini & Hochberg, 1995). Unlike a single hypothesis test, where slight nonuniformity in the p-value distribution is unlikely to change results, when testing many thousands of hypotheses slight deviations from uniformity in the p-value distribution could cause substantial changes in the set of journals that are rejected (Efron, 2007). Furthermore, it is well-known that a large number of false rejections arise when testing multiple independent hypotheses. In this section, we explain each class of hypothesis test that we perform, as well as how we account for the fact that we are testing several thousand hypotheses.

H_1: The steady-state citation distribution for papers published in a journal is a discrete lognormal distribution.

We use the χ^2 test to statistically test whether the discrete lognormal model is consistent with the data for each individual journal. The data is binned in such a way that there are at least 5 expected observations per bin based on the maximum likelihood estimate of the parameters. The assumptions necessary for the classical χ^2 test are satisfied (Taylor, 1997) when it is possible to have five or more bins with an expected count of at least five events, and we assess the significance of the χ^2 statistic using the χ^2 distribution with $b-4$ degrees of freedom, where b is the number of bins. For most journals, we have on the order of hundreds of degrees of freedom. However, for almost 1% of the journals we consider, we have fewer than 10 degrees of freedom.

Although it may appear to be questionable to fit a three-parameter model to data with fewer than 10 degrees of freedom, we note that we are making the fit to the *same* model for data with hundreds or even thousands of degrees of freedom. Because, as we will see, our three-parameter discrete lognormal model cannot be rejected, it is not appropriate to consider a different model for the small set of journals with only a few degrees of freedom.

Note that when the number of bins is smaller than five, we use a parametric Monte Carlo bootstrap approach with 10,000 bootstrap samples (Efron & Tibshirani, 1994) to assess the significance of the χ^2 statistic. Such a situation occurs, for example, when the only observed values for the number of citations that papers receive are 0, 1, and 2, allowing for a maximum of three bins.

We denote the p-value of hypothesis H_1 for journal j as p_1^j. A journal j is "rejected" if the observed χ^2 statistic is unlikely under the null hypothesis, that is, if $p_1^j < \alpha_1$, where α_1 is the *per-comparison rejection threshold*. See the Appendix for a full discussion of the statistical power of this testing procedure.

H_2: The discrete lognormal distribution describes the citation distribution to every journal.

Another hypothesis we test is whether *all* of the observed data in all 2,184 journals are consistent with the discrete lognormal model. Assuming the citation data is drawn from a discrete lognormal distribution for every journal, the distribution of p_1^j will be uniform between 0 and 1. Using a per-comparison rejection threshold of α_1 for every test, there is a probability α_1 that the test will reject the null hypothesis. This process for multiple tests is analogous to flipping a weighted coin some number of times. The actual number of rejections at α_1 for the set of 2,184 journals is compared to the number expected to occur in a binomial process. Thus, to test whether all of the journals are consistent with the discrete lognormal, we use the number of rejected journals as a test statistic and assess the significance of this statistic, p_2, using the binomial distribution, $B(2184, \alpha_1)$.

H_3: The steady-state citation distribution for papers published in journal j is consistent with the discrete lognormal model, when years are considered separately.

Although H_1 assumes that the distribution during the "steady-state" period is not changing, we also want to test whether one can reject the hypothesis that the data was drawn from a discrete lognormal distribution, but that the parameters of the distribution vary in time. To test this hypothesis, H_3, we use the procedure described for H_1, except we test the hypothesis for each year in the steady-state period separately and the model is rejected for year y if $p_{3,y}^j < \alpha_3$. Then, we assess the significance of the number of rejected *years* for journal j using the binomial distribution, $B(N_Y^j, \alpha_3)$, where N_Y^j is the number of years in the steady state for journal j. We denote the p-value of hypothesis H_3 for journal j as p_3^j.

Multiple Testing Considerations

Because we are simultaneously testing multiple hypotheses, it is necessary to account for the fact that at a given

DOI: 10.1002/asi

per-comparison rejection threshold, α_1, one expects to reject a fraction α_1 of hypotheses for which the hypothesis is actually true. Keeping the nomenclature consistent with the literature on multiple testing (Benjamini & Hochberg, 1995), we refer to these false rejections as *false discoveries*. For example, if we set $\alpha_1 = 0.05$, we expect there to be $2184 \times 0.05 \approx 109$ journals rejected, even if the discrete lognormal distributions is the "true" model for every journal.

It is clear that the chosen value of α_1 governs a trade-off between the number of false discoveries (drawn from discrete lognormal distribution, but have $p_i^j < \alpha_1$) and false negatives (not drawn from discrete lognormal distribution, but have $p_1^j \geq \alpha_1$). As we are interested not only in how many journals can be rejected but also in identifying common traits of journals that are rejected, it is more important to discover a set of journals, which we can be confident mostly comprises true discoveries. Therefore, using the algorithm described in Benjamini and Hochberg (1995), we set $\alpha_1 = \alpha_{FDR} = 0.0007$, which controls the false discovery rate (FDR) at a level of 0.05. Therefore, in the set of journals for which $p_1^j < \alpha_{FDR}$, we expect only a fraction $FDR = 0.05$ of the journals to be false discoveries.

Results

Figure 2 illustrates the goodness of fit of the discrete lognormal distribution for two journals, *Circulation* and *Science*. For 229 of the 2,184 journals in our study, including *Science*, we reject hypothesis H_1, that the steady-state citation distribution is discrete lognormal, at the $\alpha_1 = 0.05$ confidence level. If hypothesis H_1 is true for every journal in the set, then we expect 109 ± 20 journals to be rejected at $\alpha_1 = 0.05$. Clearly, we observe more rejections than would be expected under hypothesis H_2; thus, we can reject ($p_2 < 0.001$) that *all* the citation distributions to papers in journals in our study could have been drawn from the discrete lognormal distribution. Note, however, that even for *Science* the fit appears visually to be quite good.

One natural question to ask is whether journals that are *not* consistent with the discrete lognormal model share any traits. For example, if we hypothesize that within a subfield the number of citations is distributed lognormally, then we might expect the distribution to be a mixture of lognormal distributions for journals that span more than one subfield. Figure 3 suggests that the journals for which hypothesis H_1

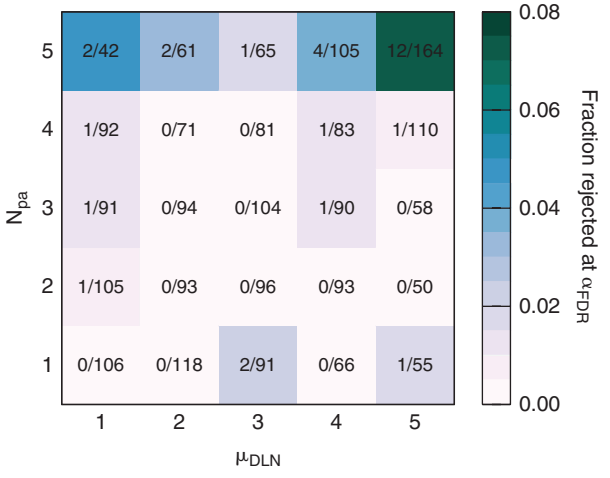

FIG. 3. Hypothesis H_1 tends to be rejected more often for high volume journals. We divide journals in our analysis by quintile according to two attributes: number of papers per annum, N_{pa}, and the best fit mean of the discrete lognormal model, μ_{DLN}. Most (70%) of the journals for which hypothesis H_1 is rejected are in the top quintile of N_{pa}. This is to be expected to some extent, since the test has more statistical power to detect even small deviations from the model for larger N (see Appendix). However, among the journals in the top quintile of N_{pa}, we find that 57% of the journals for which hypothesis H_1 is rejected are in the top quintile of μ_{DLN}. That is, large, highly-cited journals are the most likely to fail the test.

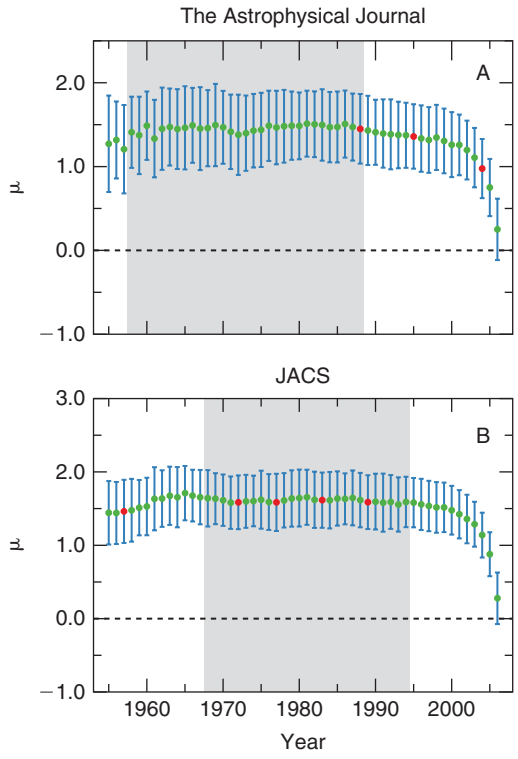

FIG. 4. Hypothesis H_3 is rejected for seven journals. In the set of journals for which hypothesis H_1 is rejected at α_{FDR}, some tests fail because the parameters of the discrete lognormal distribution actually vary slightly in time. Panel A shows the mean of the discrete lognormal distribution as a function of time for *The Astrophysical Journal* (Ap. J.). The error bars are intended to show the "width" of the distribution, or the standard deviation σ_{DLN}, as opposed to the estimated error. For *The Astrophysical Journal*, none of the individual years are inconsistent with the discrete lognormal model, $\forall y \in \{1958, 1959, \ldots 1988\}$: $p_{3,y}^{Ap.J.} > 0.05$. However, when the data from all years in the steady state period (shaded) are aggregated, $p_1^{Ap.J.}$ is low enough for hypothesis H_1 to be rejected with high confidence. In Panel B, we see that for *Journal of the American Chemical Society* (JACS) there are 4 years out of 20 for which $p_{3,y}^{JACS} < 0.05$. This number of rejections is sufficient to rejected hypothesis H_3 at $p_3 < 0.001$. Thus, the time varying mean is not sufficient to explain the deviations from the model expectations for JACS. For purposes of estimating the ultimate number of citations that papers will receive, the heuristic method for determining a "steadystate" is adequate. However, when the number of papers is large enough for the test to be very sensitive, we see that the distribution is actually a mixture of discrete lognormal distributions with a time-varying mean.

DOI: 10.1002/asi

TABLE 1. Journals for which hypothesis H_1 is rejected, ordered by p_1.

Journal	N_{pa}	Y_i	Y_f	p_1	p_3
Journal of the American Chemical Society	1267	1968	1995	0.00000	0.04
Analytical Biochemistry	395	1960	1990	0.00000	0.18
Tetrahedron Letters	1341	1960	1999	0.00000	1.00
Science	921	1988	1996	0.00000	0.00
Physical Review Letters	1860	1970	2007	0.00000	0.01
Journal of Immunological Methods	257	1972	1998	0.00000	1.00
JAMA	536	1996	1998	0.00000	0.00
Journal of Chemical Education	336	1969	2000	0.00000	0.20
Annals of the New York Academy of Sciences	1216	1980	2001	0.00000	0.28
Journal of Molecular Biology	354	1973	1994	0.00000	0.28
Gene	213	1977	1991	0.00000	0.15
Tetrahedron	598	1977	1994	0.00000	0.21
The Astrophysical Journal	766	1958	1989	0.00000	0.80
Journal of Magnetic Resonance	181	1971	1993	0.00000	0.08
JETP Letters	343	1972	2002	0.00000	0.06
Obstetrics and Gynecology	387	1980	1996	0.00000	0.19
American Journal of Physics	175	1961	2000	0.00001	0.59
Archives of Biochemistry and Biophysics	560	1982	1996	0.00002	1.00
Wildlife Research	59	1991	1998	0.00005	1.00
Journal of Organic Chemistry	1023	1969	1990	0.00008	1.00
Arthroscopy	62	1992	1993	0.00028	0.05
Journal of Applied Polymer Science	674	1989	1995	0.00032	0.26
Allergologie	73	1982	2004	0.00034	0.28
Echocardiography	87	1990	2004	0.00041	0.00
Journal of Neurochemistry	332	1958	1998	0.00042	0.28
American Journal of Human Genetics	294	1990	2002	0.00046	1.00
Cereal Chemistry	93	1955	1996	0.00048	0.00
Psychotherapy and Psychosomatics	47	1980	2004	0.00051	0.32
Archives of Dermatology	179	1968	1996	0.00059	0.41
Biochemical Society Transactions	239	1975	2007	0.00060	0.04

Note. These journals were identified by setting the false discovery rate to 0.05. That is, we expect up to 4 of these journals to be falsely rejected. N_{pa} is the average number of papers published per annum during the steady state period. Y_i and Y_f correspond to the start and end of the steady state, respectively. p_3 is the p-value for the year-by-year hypothesis testing procedure.

is rejected are primarily journals that publish many papers each year. Additionally, among the rejected journals, there is a dependence on μ_j, as there are more rejected journals in the top quintile than would be expected by chance.

The first observation could be at least partially explained by the size dependence of the statistical power of the testing procedure (see Appendix). In journals with more papers, the probability of detecting even small deviations from the discrete lognormal model increases. The second observation, however, cannot be an artifact of the statistical power of the testing procedure, because the power depends weakly on μ_j (see Appendix). However, it could be that high-impact, high-volume journals are more likely to be multidisciplinary; thus, papers published there would have heterogeneous citation properties.

Under the working hypothesis that papers published within a single subfield will have discrete lognormal citation distributions, the fact that *some* journals fail could suggest two possible explanations: (a) the distribution during the steady state is actually changing, and, thus, the distribution is a mixture of lognormal distributions due to a time-varying mean, or (b) high-volume journals are more likely to publish the research falling within distinct scientific subfields, each having different citation behaviors, and, thus, the distribution is a mixture of lognormal distributions due to heterogeneous citation properties.

The first explanation is plausible because the heuristic method described in Stringer et al. (2008) detects distributions that are statistically similar, but not necessarily statistically indistinguishable. To test if this is indeed occurring, we tested each year in the steady-state period separately using hypothesis testing procedure H_3. Figure 4 shows the citation distribution history for two journals. Table 1 lists the journals for which the discrete lognormal model fails, using the rejection threshold of α_{FDR} (see the Methods section). For 76% of the 30 journals for which the steady-state data is inconsistent with the discrete lognormal model, we found that we cannot reject the discrete lognormal model when each year is considered separately and the multiple testing procedure is used to assess the significance. Thus, it appears that among the rejected journals, the primary reason for rejection is that the mean is changing slightly in time and that the steady state distribution is not really steady "enough." Table 2 lists the journals for which hypotheses H3 is rejected. Note that multidisciplinary journals are over-represented in this set.

DOI: 10.1002/asi

TABLE 2. Journals for which hypothesis H_3 is rejected, ordered by p_3.

Journal	N_{pa}	Y_i	Y_f	p_3
Science	921	1988	1996	0.00
Cereal Chemistry	93	1955	1996	0.00
JAMA	536	1996	1998	0.00
Echocardiography	87	1990	2004	0.00
Physical Review Letters	1860	1970	2007	0.01
Biochemical Society Transactions	239	1975	2007	0.04
Journal of the American Chemical Society	1267	1968	1995	0.04

Note. These journals can be confidently rejected, even when individual years are tested separately. Note that multidisciplinary journals are over-represented in this set. We conjecture that these journals are not consistent with the lognormal model because they are a publication outlet for more than one subdiscipline.

Discussion

Our results demonstrate that for an overwhelming majority of journals, from every discipline covered by WoS, the distribution of the number of citations to papers published in that journal is consistent with a discrete lognormal model. This implies that for the logarithm of the number of citations, there is a typical value for the number of citations that a paper will ultimately receive, which depends on the journal and year in which the paper is published. For a large majority of journals, this typical value is constant in time after the initial citation accumulation period. The implications of this finding can be valuable for those wanting to compare publications based on their citation impact. Indeed, two of the primary criticisms of citation analysis methodology can be addressed by following two simple procedures: (a) waiting long enough for the journal or paper to accumulate the majority of citations that it will ultimately receive, and (b) using the logarithm of the number of citations as the quantification of impact (as in this case, our intuition about normally-distributed variables holds).

The existence of a normally distributed latent variable and a simple mapping between the value of that variable and number of citations raises the interesting prospect that each journal has a characteristic value for the "citation propensity" or "latent rate" (Burrell, 2003) of articles published therein. This citation propensity could be used to evaluate the effectiveness of journal peer review in selecting papers that will likely have high impact (Bornmann & Daniel, 2008). In addition, in modeling growing citation networks, it could be interesting and useful to explicitly account for the presence of journals and formulate models that reproduce a lognormal distribution of citations. Such models would likely be more realistic representations of the growth of the network of scientific papers.

Interestingly, our results suggest that the global citation distribution over all publications is a mixture of discrete lognormal distributions. In fact, Perline (2005) describes how a mixture of lognormal distributions can mimic a power law, which may explain why many previous studies have reported power law distributions. However, one may equally well ask why the global distribution of citations is important. Scholarly communication practices such as citation behavior, peer review type, typical peer review timescales, and others vary by field (Cole, 2000). Therefore, to avoid obfuscating one's analysis by comparing scholarly publications that are not comparable, it is important to restrict the analysis to fields in which papers have homogeneous citation properties.

References

Adam, D. (2002). Citation analysis: The counting house. Nature, 415(6873), 726–729.

Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability mass at the origin. Journal of the American Statistical Association, 50, 901–908.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate—A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B, 57(1), 289–300.

Bornmann, L., & Daniel, H. (2008). 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. Journal of the American Society for Information Science and Technology, 59(11), 1841–1852.

Boyack, K.W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. Scientometrics, 64(3), 351–374.

Burrell, Q. (2001). Stochastic modelling of the first-citation distribution. Scientometrics, 52, 3–12.

Burrell, Q. (2003). Predicting future citation behavior. Journal of the American Society for Information Science and Technology, 54(5), 372—378.

Cole, J.R. (2000). A short history of the use of citations as a measure of the impact of scientific and scholarly work. In B. Cronin & H. B. Atkins (Eds.), The Web of knowledge: A festschrift in honor of Eugene Garfield (pp. 281–300). Metford, NJ: Information Today Inc. ASIS Monograph Series.

Efron, B. (2007). Size, power and false discovery rates. Annals of Statistics, 35(4), 1351–1377.

Efron, B., & Tibshirani, R.J. (1994). An introduction to the bootstrap. Boca Raton, FL: Chapman & Hall/CRC.

Egghe, L., & Rousseau, R. (2000). Aging, obsolescence, impact, growth, and utilization: Definitions and relations. Journal of the American Society for Information Science, 51, (11), 1004–1017.

Experts still needed. (2009). Nature, 457(7225), 7–8.

Glänzel, W., & Garfield, E. (2004). The myth of delayed recognition. Scientist, 18(11), 8–8.

Krapivsky, P., & Redner, S. (2001). Organization of growing random networks. Physical Review E, 63, 066123.

Lane, J. (2009). Assessing the impact of science funding. Science, 324(5932), 1273–1275.

Lotka, A.J. (1926). The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences, 16(12), 317–323.

Nadarajah, S., & Kotz, S. (2007). Models for citation behavior. Scientometrics, 72(2), 291–305.

Not-so-deep impact. (2005). Nature, 435(7045), 1003–1004.

DOI: 10.1002/asi

Perline, R. (2005). Strong, weak and false inverse power laws. Statistical Science, 20(1), 68–88.
Press, W.H., Teukolsky, S.A., Vetterling, W.T., & Flannery, B.P. (2002). Numerical recipes in C: The art of scientific computing (2nd ed.). New York: Cambridge University Press.
Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. Proceedings of the National Academy of Science, USA, 105(45), 17268–17272.
Redner, S. (1998). How popular is your paper? an empirical study of citation distribution. European Physical Journal B, 4(2), 131–134.
Redner, S. (2005). Citation statistics from 110 years of physical review. Physics Today, 58(6), 49–54.
Shiffrin, R.M., & Borner, K. (2004). Mapping knowledge domains. Proceedings of the National Academy of Science, 101(Supl 1), 5183–5185.
Shockley, W. (1957). On the statistics of individual variations of productivity in research laboratories. Proceedings of the Institute of Radio Engineers, 45(3), 279–290.
Simkin, M., & Roychowdhury, V. (2007, Sep.). A mathematical theory of citing. Journal of the American Society for Information Science and Technology, 58(11), 1661–1673.
Solla Price, D.J.de. (1976). A general theory of bibliometric and other cumulative advantage processes. Journal of the American Society for Information Science, 27, 292–306.
Stewart, J. (1994). The poission-lognormal model for bibliometric scientometrics distributions. Information Processing & Management, 30(2), 239–251.
Stringer, M., Sales-Pardo, M., & Amaral, L. (2008). Effectiveness of journal ranking schemes as a tool for locating information. PLoS ONE, 3(2), e1683.
Taylor, J.R. (1997). An introduction to error analysis. Sausalito, CA: University Science Books.
Tomlin, S. (2005). The expanding electronic universe. Nature, 438(7068), 547–555.
Wouters, P. (1999). The citation culture. Unpublished doctoral dissertation, University of Amsterdam.

Appendix

The probability of having a p-value in the rejection region given that the null model is actually false is the power of the statistical hypothesis test. The power of the test depends on the deviation from the null hypothesis, which is unknown for observed data. In addition, the power depends on the number of observed samples. To investigate the power of our testing procedure, we performed a Monte-Carlo simulation experiment of plausible deviations from the null hypothesis. Figure A1 summarizes the results.

We hypothesize that if a journal is multidisciplinary, then it may be the case that the distribution is a mix of discrete lognormal distributions. For simplicity, we consider the case when a journal covers two disciplines with different mean citation propensities and a model in which the latent variable distribution is a mixture of two normal distributions. A fraction m of the citation propensities are drawn from the distribution $N(\mu_1, \sigma)$ and the remaining fraction $1 - m$ is drawn from the distribution $N(\mu_2, \sigma)$. The means of the component distributions are separated by $\Delta\mu$. The midpoint between the two component distributions, μ_C is held constant at a value of 1.0. The value of σ is fixed at a 0.5. The number of samples, N_{sample} is 400, 2,000, and 10,000. The difference in means, $\Delta\mu$, is $1\sigma = 0.5$, $2\sigma = 1.0$, and $3\sigma = 1.5$. The fraction of data drawn from the mixture component with the lower mean, m, is varied from 0.25, 0.5, and 0.75. For each set of parameters, the following procedure is repeated 10000 times:

1. Draw N_{sample} numbers from the normal mixture distribution.
2. Estimate the parameters of the discrete lognormal model.
3. Use the χ^2 as described before to obtain the p-value.
4. Reject the discrete lognormal model if $p \leq \alpha$.

Because we know that none of the synthetic datasets are actually generated from the discrete lognormal model, there can be no false negatives, the fraction of times that the discrete lognormal model is rejected is an estimate of the power of the test.

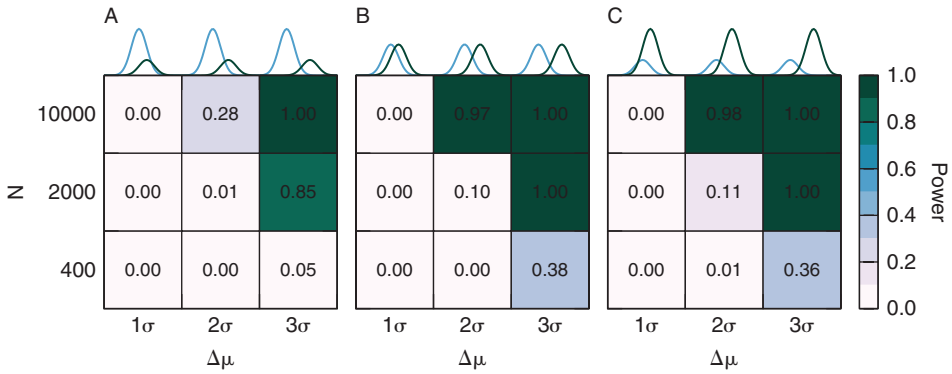

FIG. A1. Power analysis of our testing procedure. We draw data from a mixture of two normal distributions with different means but identical variance. 25% (A), 50% (B), and 75% (C) of the points in these samples were drawn from the normal with the largest mean. N_{sample} is the number of data points in each sample considered and $\Delta\mu$ is the difference in the means of the distributions. The power of the testing procedure depends on both the specific nature of the deviation from the null hypothesis and the number of points in the dataset. It is clear that a distribution that results from a mixture of lognormal distributions that are separated by only one standard deviation are essentially impossible to detect.

DOI: 10.1002/asi