

# Empirical Patterns in Google Scholar Citation Counts

Peter T. Breuer

Department of Computer Science  
University of Birmingham  
Edgbaston, Birmingham B15 2TT, UK  
Email: ptb@cs.bham.ac.uk

Jonathan P. Bowen

School of Computing, Telecomms and Networking  
Birmingham City University  
Millennium Point, Curzon Street  
Birmingham B4 7XG, UK  
Email: jonathan.bowen@bcu.ac.uk  
URL: <http://www.jpbowen.com>

**Abstract**—Scholarly impact may be metricized using an author’s total number of citations as a stand-in for real worth, but this measure varies in applicability between disciplines. The detail of the number of citations per publication is nowadays mapped in much more detail on the Web, exposing certain empirical patterns. This paper explores those patterns, using the citation data from Google Scholar for a number of authors.

## I. BACKGROUND

The speed of transmission and the quantity of knowledge available to researchers has accelerated dramatically in recent decades with the advent of the Internet and the World Wide Web. Whereas, previously, academic papers were really published only on paper, in journals and books, now they can be and often are communicated ‘online’. That has led to this body of information on academic activity becoming ever more comprehensively indexed.

Google, as well as ubiquitously indexing all of the Web, provides an index of academic publications in particular through its *Google Scholar* website (<http://scholar.google.com>) and also provides access to books through the *Google Books* facility (<http://books.google.com>). It thus has a very complete and continuously updated collection of academic data available, and is arguably currently the leading facility of that kind. Microsoft Academic Search (<http://academic.research.microsoft.com>) provides a competing database of academic publications online, started at Microsoft’s Beijing research laboratory. While it is not as complete or up to date as Google Scholar, it does provide better visualization facilities.

Google Scholar furnishes individual authors with a personalizable page that presents a list of their own publications and links to the publications that cite them, with counts of the number of citations per publication. The page is generated automatically, but it must be corrected by hand by the author in order to obtain an accurate record. Google’s automated scanning of online publications works well for popularly cited works, because multiple examples in different texts enable Google’s automata to learn to recognize the citation despite differences in spelling and presentation.

For most productive academic authors, however, there is a ‘long tail’ in the automatically generated data that consists of those publications with few or no citations, for which Google’s data can be inaccurate and may well need correction. Authors with common names may find publications by other authors with the same or similar names wrongly assigned to their page, for example. Google Scholar also confuses publications that have the superficial appearance of papers (e.g., programme committee information for conferences) with real papers, and

such entries need to be pruned. Conversely, publications that are not represented on the Web at all will not be located by Google, and must be added in by hand. Those publications that have appeared on the Web in slightly different forms ought also to be merged into a single base entry on the author’s page. An author’s corrections are not cross-checked before appearing online in Google Scholar, while corrections made in Microsoft Academic Search are checked before incorporation, with a delay until the submitted updates appear.

In spite of the more problematic aspects, a Google Scholar page provides a comprehensive opportunity for administrators to garner raw statistics on an academic’s output, which may affect prospects for promotion and tenure, and it is therefore likely that a good proportion of academics are aware of their own page’s existence and have paid some attention to ensuring that it is reasonably accurate, as well as monitoring it to see how the numbers grow with time. The authors of this paper are among them; we have wondered why the data on our pages looks the way it does and if there is some underlying pattern to it that we should be seeing. This paper points out some of the patterns we have empirically observed and puts forward a theory as to their causes.

## II. CITATION METRICS

There are a number of metrics in use that have the common aim of measuring the stature of an academic researcher in their field. The simplest is an author’s total citation count, but it has a number of drawbacks. Authors often have a large number of publications with relatively low numbers of citations that have had comparatively little influence on their field, so why account them? Most researchers of influence have only a small set of key publications that have been highly cited by their peers. For example, Alan Turing [2] had three key publications with thousands of citations, each of which have led to the foundation of important areas of computer science, whereas some of his works have never been cited (and we will resist the temptation to skew the statistics by citing them here; the reader should check Alan Turing’s Google Scholar page). Normally, it is the leading papers by an author that are accepted as being significant, and the total citations count obscures the nature of their contribution. The more highly regarded metrics weigh the ‘top end’ of an author’s output more heavily. The maximal citation count alone is better than the total for that purpose, but Alan Turing would have his second and third counts, which are only insignificantly different from the first, not taken into consideration by that method.

Google Scholar derives the author’s *h-index* [7] and *i<sub>10</sub>-index*, which are popular indications of respectively the depth

and breadth of an author’s impact, especially in the field of computer science. The  $i_{10}$ -index counts the number of the author’s works that have received at least 10 citations. Looking at the graph of citation counts per publication arranged in decreasing order left-to-right, it is the distance out from the left hand side at which the graph falls below 10 in height. All other things being equal, the higher the total number of citations in the graph, the greater will be the  $i_{10}$ -index – that is, assuming that all researcher’s graphs have roughly the same shape. But are they the same shape for different researchers?

The h-index is that value of  $n$  in the  $i_n$  number that approximately satisfies  $i_n = n$ . It is the length of the side of the biggest square that can be placed under the graph of citation counts per publication arranged in decreasing order left-to-right. So it measures how ‘fat’ the graph is near the mid-point, in a direction out at  $45^\circ$  to the axes from the origin. How might this be expected to change with respect to total number of citations, for example?

These measures need to be treated with caution, since different fields have significant variation in their patterns of publication [12]. For example, computer scientists tend to have a much lower number of co-authors than physicists, whose co-authors may play a very small role in a highly cited paper. Radicchi et al in [10] suggest that all citation counts ought to be normalised with respect to the average citation count per article in their field, and they note that such normalized counts seem to follow the same distribution independent of field. Nevertheless, the individual numbers for an author’s output are interesting, and we can observe certain empirical relations between them, which we will elaborate below.

### III. PATTERNS IN CITATIONS

Is there a pattern to the graph of citation numbers per article for a given author? On the default Google Scholar page, the author’s articles are listed with the citation numbers decreasing down the page. Evidently, there are more entries with, say, 4 citations than there are with 40, but if one erases the numbers and replaces them with a plain graph, can one tell how far down the page one is by the way the graph looks?

The answer is a qualified no – the graph of citation counts in decreasing order for a given author in Google Scholar appears to be ‘scale invariant’: one part of the graph looks just the same as another part, scaled up or down by a factor.

Evidence of that comes from the result of a test of Benford’s law [1] on scale-invariant distributions. Benford’s law says that the number of individual article citation counts that start with the digit 1 should be greater than the number that start with 2, and so on. The violation of that law for published Iranian election counts was taken as an indication that the results had been tampered with [11]. For the two authors of this paper, the frequencies for the first digits of the citation numbers of their articles as listed in Google Scholar are shown in Table I. They are consistent with the expectation for scale-invariant distributions – a decrease in the observed frequencies from the smaller digits to the larger digits, ideally down from about 30% of the total for digit 1 to about 5% of the total for digit 9.

With that evidence for scale-invariance to hand, the question is what kind of scale-invariant distribution is it. We believe that the Google Scholar citation counts for articles, laid out in

decreasing order from the article with highest citation count to the lowest approximately follows an exponential power-law:

$$c_n \approx c_0 e^{-P\sqrt{n}} \quad (1)$$

where  $c_n$  is the citation count of the  $n$ th article in decreasing order of citation count, and  $c_0$  is the citation count of the most cited article. The factor  $c_0$  in (1) is chosen so that the approximation becomes exact at  $n = 0$ .

#### A. Estimating the multiplier

Achieving insight into the Google Scholar data for an author entails getting a good estimate for the multiplier  $P$  in (1). Below we suggest four increasingly fit-for-purpose calculations for  $P$ . First, putting  $n = 1$  in (1) gives:

$$P \approx -\ln(c_1/c_0) \quad (2)$$

This is the natural logarithm of the ratio of the citation counts of the most cited two articles. This calculation emphasises that  $P$  is related to the initial slope of the graph. Bigger  $P$  means a steeper initial slope and the ‘punchier’ is the author’s best compared to the rest. Whether that is good or bad shall be left to the reader to decide.

Another estimate comes from putting  $n = i_1$  in (1), where  $i_1$  is the number of cited articles, the least  $n$  such that  $c_n = 0$ :

$$\begin{aligned} i_1 &\triangleq \#\{n \mid c_n \geq 1\} \\ &= 1 + \max\{n \mid c_n \geq 1\} \\ &= \min\{n \mid c_n = 0\} \end{aligned}$$

and thus  $1 \approx c_0 e^{-P\sqrt{i_1}}$ , giving rise to

$$P \approx \frac{\ln c_0}{\sqrt{i_1}} \quad (3)$$

This can be interpreted as the ‘sharpness’ of the roughly triangular shape formed by the graph of the  $\ln c_n$  against  $\sqrt{n}$ . If an author has a highly cited article, this estimate is larger. But if an author has many hardly-cited articles, the estimate is lower. An author can trade off a few ‘duds’ against an increment in the order of magnitude of the most cited article. Completely uncited works do not impact this measure at all.

For the first author of this paper, a good value of  $P$  is empirically about 0.5. The citations data and the approximating curve (1) for  $P = 0.5$  are shown together in the left hand diagram in Table II. The correspondence is visually excellent.

There is further evidence for the quality of the approximation (1) in the right hand diagram of Table II, which shows the same plot in log-log format, with  $\ln(-\ln \frac{c_n}{c_0})$  being plotted against  $\ln n$ . The approximating curve is transformed by the logarithmic scaling into the straight line  $\ln P + 0.5 \ln n$ , and  $\ln P$  is where the line crosses the y-axis, near  $-0.7 = \ln 0.497$ .

The estimate (2) is  $P = 0.4$  (0.401), and (3) gives  $P = 0.5$  (0.497).

Finally, here is a holistic estimate for  $P$  that takes into account input from all the data points and which lies between (2) and (3). It is based on

$$\sum_{n=0}^{\infty} c_n \approx 2 \frac{c_0}{P^2}$$

TABLE I. FREQUENCIES (y-AXIS) OF FIRST DIGITS (x-AXIS) IN THE TWO AUTHORS GOOGLE SCHOLAR ARTICLE CITATIONS NUMBERS.

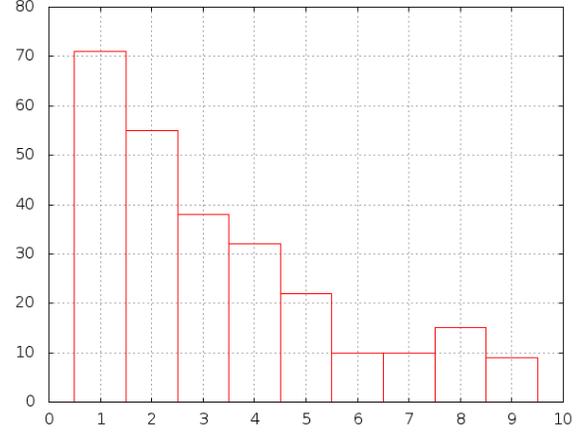
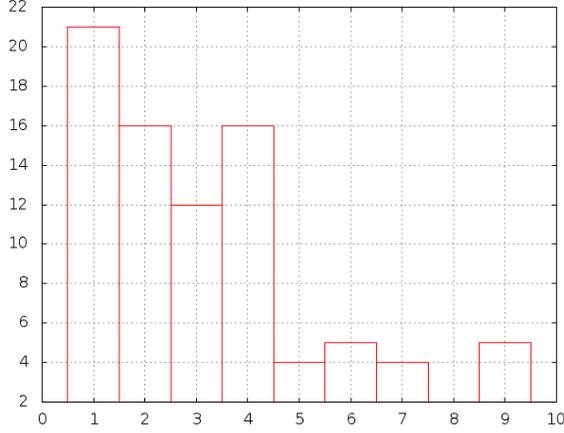
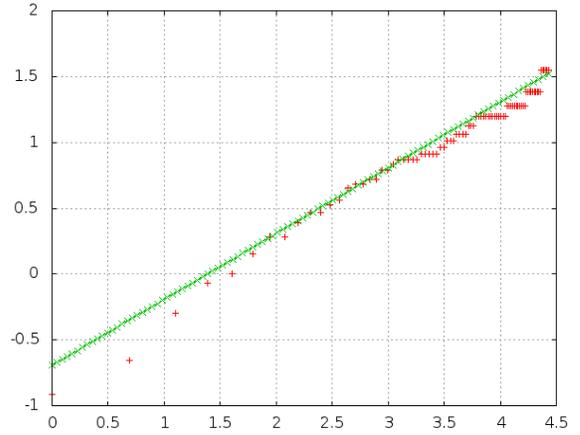
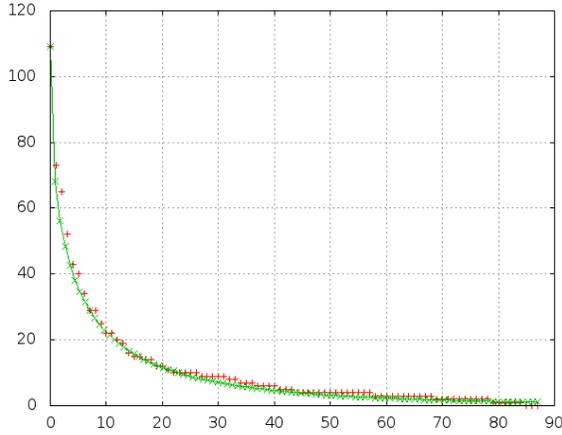


TABLE II. LEFT: GOOGLE SCHOLAR CITATION COUNTS FOR THE FIRST AUTHOR, AGAINST THE APPROXIMATION  $c_n = c_0 e^{-P\sqrt{n}}$  WITH  $P = 0.5$ . RIGHT: THE SAME DATA IS SHOWN ON A LOG-LOG GRAPH, WITH  $\ln(-\ln \frac{c_n}{c_0})$  AGAINST  $\ln n$ .



where the right hand side is the area under the curve  $c_0 e^{-P\sqrt{x}}$  from zero to infinity. The left hand side is the total number of citations for the author, which is also a rough measure of the same area on the plot. We will call this total  $S$  from now on:

$$S \triangleq \sum_{n=0}^{\infty} c_n$$

Thus  $S \approx 2c_0/P^2$ , giving the estimate:

$$P \approx \sqrt{2c_0/S} \quad (4)$$

For the first author of this paper, (4) yields  $P = 0.47$ , between the 0.5 posited from the acuteness of the log/root triangle formed by the citations graph, and the 0.40 from the log of the ratio of the top two citation numbers. It is a balanced estimator over the whole set of data, but in consequence the first few citation numbers (the most cited on down) are not so well approximated by it. The effect is visible in the log-log plot (Table II), where the first few data points appear off and below the approximating curve although the tail is well approximated. On the plot with unscaled axes, however, the deviation is hardly noticeable.

The fourth estimate of  $P$  comes from supposing the value of  $\ln P$  is where a best-fit straight-line approximation on the log-log graph of citations crosses the y-axis. Formalised in terms of the covariance and averages of the log-log data, it is

$$\ln P \approx \text{av}(\ln(\ln \frac{c_0}{c_n})) - \text{av}(\ln n) \text{cov}(\ln(\ln \frac{c_0}{c_n}), \ln n) \quad (5)$$

where the uncited papers and the very top cited paper are left out of the reckoning here. The arithmetic averages of the logarithms are the logarithms of the geometric means.

### B. Adjusting the shape of the curve for different authors

We do not yet have an underlying rationale for the term  $\sqrt{n}$  in the empirically observed formula (1). It seems not to be quite right for some authors, and in general we would like to suppose that the term is  $n^A$  for some constant  $A$  that just happens to be approximately  $A = 0.5$  in the case of the first author of this paper. The more general approximation is:

$$c_n \approx c_0 e^{-Pn^A} \quad (6)$$

and (1) is (6) with  $A = 0.5$ . A log-log plot like that on the right in Table II allows  $A$  to be estimated by the slope of the

approximating straight line, and placing the line by eye on the plot is a good practical means of positing a value for  $A$ .

For Alan Turing's citation data as shown on Google Scholar, we obtain a better approximation with  $A = 0.4$  than with  $A = 0.5$ . The approximation with  $A = 0.5$  is shown in green in Table III. The approximating curve is high in mid-range, and a little low farther up-range. That is not too surprising, given the abnormality of Turing's data. He has three most cited papers of roughly the same order, and then a fourth and more papers an order of magnitude less cited (but still enormously highly cited by most standards). We could never hope to capture three nearly equal top papers with the kind of approximation in (6) and  $|A| < 1$  because of the sharp peak that approximation produces at zero. Turing's data is more like what one would expect from three contributors, or three equal careers (indeed, one of the top three papers is in mathematical biology). Still the log-log curve shows clearly that 0.5 is too steep a slope for the straight line approximation after the first few points. We do need an approximating term more like  $2.5\sqrt{n}$  than  $\sqrt{n}$  for the tail.

Plotting Alan Turing's data against  $A = 0.4$  gives the blue lines in Table III. The log-log graph looks perfect.

How does one estimate  $P$  numerically in the general case?

The 'area under the curve' argument on the unscaled graph gives the following approximation when  $A$  is the reciprocal of an integer,  $A = \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots$  (but not, yet, for  $A = 0.40$ ):

$$P \approx (c_0 (A^{-1})! / S)^A$$

For other values of  $A$ , we need to replace the factorial expression using Euler's gamma function:

$$P \approx (c_0 \Gamma(1 + A^{-1}) / S)^A \quad (7)$$

Varying  $A$  then allows approximations to be fine-tuned.

For the second author of this paper,  $A = 0.4$  also appears to be better than  $A = 0.5$ . The plots for  $A = 0.4$  are shown in Table IV, and  $P = 0.5$  is approximately right for it by eye. In this case  $\Gamma(1 + A^{-1}) = \Gamma(\frac{5}{4}) = \frac{15}{8}\sqrt{\pi}$ , so (7) becomes

$$P \approx \left( \frac{15c_0}{8S} \sqrt{\pi} \right)^{0.400}$$

which is 0.520 because  $c_0/S = 0.0587$  for this author. The number is in good agreement with the visually fitted value of  $P = 0.5$ .

### C. Classical citation-based measures of worth

The  $i_{10}$ ,  $i_{20}$  measures promoted on Google Scholar and elsewhere are defined by  $c_{i_k-1} \geq k > c_{i_k}$ , so:

$$\begin{aligned} 10 &\approx c_{i_{10}} \approx c_0 e^{-P(i_{10})^A} \\ 20 &\approx c_{i_{20}} \approx c_0 e^{-P(i_{20})^A} \end{aligned}$$

Thus

$$\begin{aligned} \ln(10/c_0) &\approx -P(i_{10})^A \\ \ln(20/c_0) &\approx -P(i_{20})^A \end{aligned}$$

and

$$\ln(10/c_0) / \ln(20/c_0) \approx (i_{10}/i_{20})^A$$

or

$$i_{10}/i_{20} \approx A \sqrt{\ln(c_0/10) / \ln(c_0/20)} \quad (8)$$

For the first author of this paper and  $A = 0.5$ , the predicted ratio (8) is 1.98, against the real ratio  $i_{10}/i_{20} = 2.08$ . The estimate is very close. For the second author of this paper and  $A = 0.4$ , the predicted ratio (8) is 1.75 and  $i_{10}/i_{20} = 1.69$  in reality, again very close. For Alan Turing, with  $A = 0.4$  the predicted ratio is (8) is 1.32 and  $i_{10}/i_{20} = 1.44$  in reality, somewhat less close, but Turing's numbers are extraordinary. It is striking, however, that the prediction adjusts to approximately match the author in every case, despite their differences.

We can derive a relationship between  $i_{10}$  and the h-index, using  $i_h \approx h$  and replacing both  $i_{20}$  and 20 by  $h$  in (8). Then

$$(c_0/h)^A \ln(c_0/h) \approx (c_0/i_{10})^A \ln(c_0/10) \quad (9)$$

For the first author and  $A = 0.5$ , the function on the left is  $(109/h)^{1/2} \ln(109/h)$ , and the constant on the right is 4.8, with solution  $h \approx 16.67$ . The truth is that the author's h-index is 15 – but it is only one citation away from 16.

The number on the right in (9) is fairly constant across a range of i-indices for a given author. For the first author of this paper, it is in the range 4 to 5 up to about  $i_{50}$ . For the second author of this paper, it is in the range 5 to 6 up to about  $i_{100}$ . A larger number means that the ratio  $c_0/h$  is larger, and the h-index is a smaller fraction of the peak ('most-cited') number.

## IV. UNDERLYING CAUSES

How can we explain these observations?

Plotting the number of articles for which the number of citations  $C$  falls in  $x \leq \ln C / \ln c_0 < x + dx$  against  $x = \ln C / \ln c_0$  for  $C \geq 1$  gives a curve with mean  $\mu$  and standard deviation  $\sigma$ . That is,  $(\ln C / \ln c_0 - \mu) / \sigma$  looks like a random variable with mean 0 and standard deviation 1. Moreover, in the data sets we have looked at, the mean  $\mu$  and standard deviation  $\sigma$  are nearly the same (approximately 0.2) in every case. For the first and second authors, and Alan Turing, the mean and standard deviation pairs  $(\mu, \sigma)$  are respectively (0.251331, 0.216663), (0.217722, 0.211669), (0.190153, 0.196518). Say:

$$\sigma = \mu = \lambda$$

We may take  $\ln C$  to be normally distributed with mean  $\mu = \lambda \ln c_0$  and standard deviation  $\sigma = \lambda \ln c_0$ . By the standard statistics of log-normal distributions, one expects citations  $C$  to have mean  $m$  and standard deviation  $s$  where

$$m = c_0^{2\lambda} \quad s = c_0^{2\lambda} \sqrt{e^{(\lambda \ln c_0)^2} - 1}$$

which allows the parameter  $\lambda$  to be conveniently estimated from  $s/m$ . For the first and second authors, and Alan Turing,  $\lambda$  estimated this way is 0.264, 0.250, 0.209 respectively.

We have generated sets of fake citation data using a normally distributed random variable for  $\ln C / \ln c_0$  with both mean and standard deviation equal to  $\lambda$ . That is,  $C = e^{(\lambda + \lambda X) \ln c_0}$  where  $X$  is a normally distributed random variable with mean 0 and standard deviation 1. The generated data looks like a real citations count list, and ordering it in descending order  $c_0, c_1, \dots$  and plotting it in log-log as in Table II-IV (with  $\ln(-\ln(c_n/c_0))$  against  $\ln n$ ) shows that a straight line approximation is appropriate. The slope of the line

TABLE III. LEFT: CITATION COUNTS FOR ALAN TURING, AGAINST THE APPROXIMATION  $c_n = c_0 e^{-0.95n^{0.5}}$  (GREEN), AND  $c_0 e^{-1.50n^{0.4}}$  (BLUE). RIGHT: THE SAME DATA ON A LOG-LOG GRAPH, WITH  $\ln(-\ln \frac{c_n}{c_0})$  AGAINST  $\ln n$ .

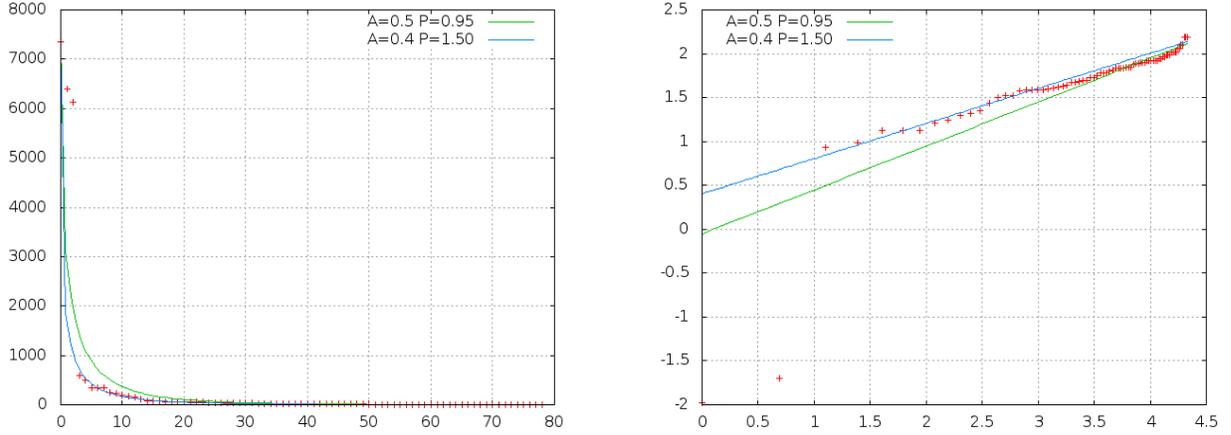
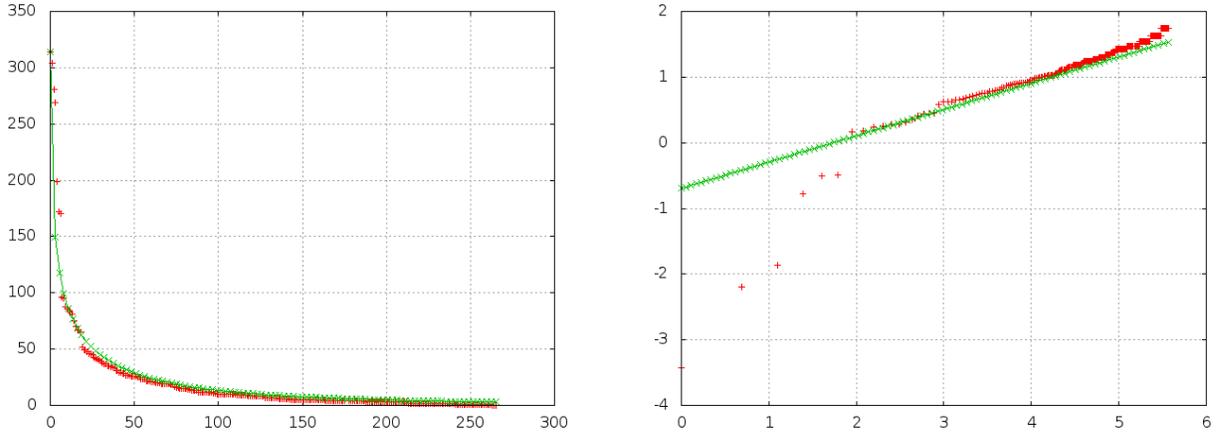


TABLE IV. LEFT: CITATION COUNTS FOR THE SECOND AUTHOR, AGAINST THE APPROXIMATION  $c_n = c_0 e^{-Pn^{0.4}}$  WITH  $P = 0.5$ . RIGHT: THE SAME DATA SHOWN ON A LOG-LOG GRAPH, WITH  $\ln(-\ln \frac{c_n}{c_0})$  AGAINST  $\ln n$ .



is the  $A$  of the approximation (6), and the slopes manifestly cluster around  $A = 0.4$ , but can vary between 0.2 and 0.6.

Placing the best fit line via least squares minimisation of the errors on the log-log plots (excluding the first citation number) gives the slope as the covariance between x- and y-ordinates of the data points. That gives the following estimates of slope  $A$  for  $N = 200$  datapoints. The average is taken over 100 generated datasets for each value of  $\lambda$  listed:

$\lambda$	slope $A$	standard deviation
0.2	0.396119	0.075180
0.25	0.405471	0.077208
0.3	0.411451	0.080782
0.35	0.384616	0.078864
0.4	0.400892	0.087734
0.45	0.413907	0.090714

This is empirical support for the approximations with power exponents  $A=0.4$  (exceptionally 0.5, in the case of Table II)

as in Tables II-IV. The value of  $\lambda$  has insignificant effect.

This value of  $A$  arises naturally. It is the slope of the best-fit line to the logarithm of normally distributed data with equal mean and standard deviation that has been ranked in decreasing order against the logarithm of the ranking position.<sup>1</sup> But the precise value depends on the number of datapoints  $N$  as follows:

<sup>1</sup> The slope is that of  $\ln(x_0 - x_n)$  against  $\ln n$  for a standard normal variable  $X$ , with the  $N$  observations arranged in decreasing order  $x_0, x_1, \dots$ . Theory says that the  $x_n$  are positioned about where the quantiles function  $q(\rho) = x \Leftrightarrow \text{prob}(X < x) = \rho$  says they should be for  $\rho = (n + 0.5)/N$ , at  $x_n \approx q((n + 0.5)/N)$ . In particular, the cumulative density function for the maximum of the  $N$  observations,  $\text{prob}(x_0 < x)$ , is  $(\text{prob}(X < x))^N$ , the  $N$ th power of the cumulative density function of an individual observation, and the position of the maximum observation is expected to be  $x_0 \approx x : \text{prob}(X > x) = 1/(2N)$ ; i.e.,  $x_0 \approx q(1 - 1/(2N))$ . For the normally distributed standard variable  $X$ , that is  $x_0 \approx \frac{1}{2}\sqrt{2 \ln N}$  asymptotically, applying classical mathematical analysis to the integral that defines  $q$ . The minimum is expected to be the same distance in the other direction. Thus the slope  $\ln(x_0 - x_{N-1})/\ln N$  is approximated by  $\ln \sqrt{2 \ln N}/\ln N$ , or  $\ln(\ln N^2)/\ln N^2$ .

$N$	slope $A$	standard deviation
100	0.463705	0.105916
1000	0.363892	0.052899
10000	0.279002	0.032274
100000	0.233939	0.017043
1000000	0.204701	0.012906

The slope  $A$  slowly decreases to zero with increasing  $N$ . The measurements in the table above vary from 0.95 to 0.85 of a predicted bounding asymptote  $\ln(\ln N^2)/\ln N^2$ .

The approximation (6) is compatible with observations from a log-normal distribution. The rate  $A$  in the exponent varies slightly according to the number of an author's publications, but is otherwise stable across authors. The number  $P$  determines how exceptional are the most cited articles with respect to the body of work of an author. For the authors of this paper,  $P = 0.5$  is about right. For Alan Turing  $P = 1.5$  is indicated, highlighting the extra significance of his three top papers relative to the rest of his (also highly significant) work.

The numbers  $1 - i_n/N$  provide a direct measurement of the cumulative density function  $\text{prob}(C < n)$  for the random variable  $C$  underlying the citations counts. The derivative is the probability density function. The logarithm of citation count ( $\ln C$ ; the x-axis of the density function) looks distributed like a Poisson distribution (the y-axis). A normal distribution with equal mean and standard deviation is a fair approximation to Poisson, and is what we have used in our analysis.

A Poisson distribution represents low probability events (citations!) accruing in several equal sized slots over time. After a while, most slots have the average number of events in, while a very few have none, and a very few have a large number of events in. The situation here is that instead of seeing, say,  $k$  slots with  $c$  events as expected according to a Poisson distribution, we are seeing  $k$  articles with  $\ln c$  citations in. We do not have good insight into that from the publications and citations point of view. Perhaps it means that the average 'intrinsic worth' of an article is distributed by a Poisson process, but that articles accumulate citations according to the exponential of their worth. Citation begets citations, in other words.

Radici et al report in [10] that citation counts relative to the average count in a field follow (the same) log-normal distribution irrespective of field. We also see log-normal distribution, but within a single author's output, so perhaps the results of [10] apply when one considers an author as defining their own academic field of study. We normalize with respect to the maximal citation count, and see equal mean and standard deviation (in logarithm), while Radici et al normalize with respect to the mean citation count and see mean equal and opposite in sign to twice the variance (in logarithm). We believe these relations are reflections of the same underlying reality. While Radici et al sought to quantify an article's worth irrespective of the field it is published in, we have reduced the question of an individual author's impact to three parameters,  $c_0$ ,  $A$  and  $P$ , which predict the curve of citation counts. How these parameters are distributed across and within academic fields remains to be discovered.

## V. CONCLUSION

This paper has noted a mathematical pattern with respect to citation counts for publications of academic authors. When

the citation counts per article are laid out in decreasing order, they follow the law

$$c_n \approx c_0 e^{-Pn^A}$$

for an appropriate multiplier  $P$  and rate  $A$  fitted to an individual author. In practice  $A$  ranges from 0.4 to 0.5, and is lower for higher publication count  $N$ , decreasing as  $\ln(\ln N^2)/\ln N^2$ . The pattern is compatible with observations of a log-normal random variable, the exponential of a normal random variable with equal mean and standard deviation.

Recognizing these empirical patterns and modelling them allows more meaningful metrics to be developed.

Further patterns could be explored among, for example, temporal information based on the year of publication and citation of papers [8]. More visualization would also be possible [3], [4], [6], [9]. Communities of authors [5], including their development and demise, could also be investigated for patterning anomalies.

*Acknowledgements:* Jonathan Bowen is grateful for financial support from Museophile Limited. Google Scholar was used to provide publication data for this paper.

Peter Breuer is grateful to Richard Gill of the University of Leiden for conversations that laid bare the statistical analysis.

## REFERENCES

- [1] F. Benford. The law of anomalous numbers. *Proceedings of the American Philosophical Society*, 78(4):551–572, March 1938. JSTOR 984802.
- [2] J. P. Bowen. Alan Turing. In W. A. Robinson, editor, *The Scientists: An Epic of Discovery*, pages 270–275. Thames and Hudson, 2012.
- [3] J. P. Bowen. Online communities: Visualization and formalization. In *Cyberpatterns 2013: Second International Workshop on Cyberpatterns – Unifying Design Patterns with Security, Attack and Forensic Patterns, Abingdon, UK, 8–9 July 2013*, 2013. arXiv:1307.6360 [cs.GR].
- [4] J. P. Bowen. A relational approach to an algebraic community: From Paul Erdős to He Jifeng. In Z. Liu, J. C. P. Woodcock, and H. Zhu, editors, *Theories of Programming and Formal Methods*, number 8051 in Lecture Notes in Computer Science, pages 54–66. Springer, 2013.
- [5] J. P. Bowen and S. Reeves. From a Community of Practice to a Body of Knowledge a case study of the formal methods community. In M. Butler and W. Schulte, editors, *17th International Symposium on Formal Methods (FM 2011)*, number 6664 in Lecture Notes in Computer Science, pages 308–322. Springer, 2011.
- [6] J. P. Bowen and R. J. Wilson. Visualising virtual communities: From Erdős to the arts. In S. Dunn, J. P. Bowen, and K. Ng, editors, *EVA London 2012 Conference Proceedings*, Electronic Workshops in Computing (eWiC), pages 238–244. British Computer Society, 2012. arXiv:1207.3420v1.
- [7] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, November 2005. arXiv:physics/0508025.
- [8] D. W. Johnston, M. Piatti, and B. Torgler. Citation success over time: Theory or empirics? *Scientometrics*, 95(3):1023–1029, June 2013.
- [9] M. J. Nelson. Visualization of citation patterns of some Canadian journals. *Scientometrics*, 67(2):279–289, 2006.
- [10] F. Radicchi, S. Fortunato, and C. Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America*, 105(45):17268–17272, November 2008.
- [11] B. F. Roukema. A first-digit anomaly in the 2009 Iranian presidential election. *Journal of Applied Statistics*, 41(1):164–199, 2014. arXiv:0906.2789.
- [12] J. B. Slyder, B. R. Stein, B. S. Sams, D. M. Walker, B. J. Beale, J. J. Feldhaus, and C. A. Copenheaver. Citation pattern and lifespan: A comparison of discipline, institution, and individual. *Scientometrics*, 89(3):955–966, 2011.