Library and Information Science Research 39 (2017) 180–188

Contents lists available at ScienceDirect

Library and Information Science Research

journal homepage: www.elsevier.com/locate/lisres

Measures of greatness: A Lotkaian approach to literary authors using OCLC WorldCat

Alon Friedman[a],[*], Jay H. Bernstein[b]

[a] *School of Information, University of South Florida, 4202 E. Fowler Ave, CIS 1040, Tampa, FL 3620, United States*
[b] *Robert J. Kibbee Library, Kingsborough Community College, CUNY, 2001 Oriental Boulevard, Brooklyn, NY 11235, United States*

ABSTRACT

This study examines the productivity, eminence, and impact of literary authors using Lotka's law, a bibliometric approach developed for studying the published output of scientists. Data on literary authors were drawn from two recent surveys that identified and ranked authors who had made the greatest contributions to world literature. Data on the number of records of works by and about selected authors were drawn from OCLC WorldCat in 2007 and 2014. Findings show that the distribution of literary authors followed a pattern consistent with Lotka's law and show that these studies enable one to empirically test subjective rankings of eminent authors. Future examination of distribution of author productivity might include studies based on language, location, and culture.

1. Introduction

Bibliometrics is often defined as the statistical analysis of data about the publication and citation of works by a specific author or publisher, commonly focusing on citations of scientific research outputs, that is, how many times research publications are cited. Research in bibliometrics has developed laws explaining not only the impact of authors within scientific fields, but also the structure of that impact. Traditionally, studies have measured scientific citations found in academic journals in a discipline to examine characteristics such as gender, institutional affiliation, productivity ranking, and format. Such an approach, though appropriate for examining how scientific disciplines develop through the productivity of individual scientific researchers, raises the question of how to measure the impact of creative writing or literature.

Educators and experts in literature have attempted to delineate a common measurement of literary works, analyzing book reviews and book citation indexes, even using the Goodreads software application, to better understand the evolution of literature. However, these approaches do not sufficiently take into account the particular ways that literature can be influential.

The notion of literary output and reputation are easily grasped on an intuitive level, but seem difficult to measure. How can the relative eminence of two literary authors be compared? Can bibliometric laws or statistical formulae contribute to how literature is understood in the same way they do for scientific publications? This study seeks to develop a technique for answering these questions by introducing a bibliometric method that measures the fame or bibliographical impact of literary authors. This type of investigation is crucial to advancing bibliometric study of library works found in OCLC WorldCat.

2. Problem statement

This study introduces an innovative approach to measuring author impact and eminence that is relevant to literature and humanities disciplines. Its approach is bibliometric to the extent that it analyzes countable manifestations of recorded information. However, its materials are not citations of articles, the standard in bibliometric studies, but bibliographic records of works related to authors by authorship, subject matter, or both. This study critically examines the results and scoring used by other researchers who have developed techniques for ranking literary authors. Analysis is based on data collected in 2007 and 2014 from OCLC WorldCat,[1] an international bibliographic database of items cataloged in libraries around the world. Between 2007 and 2014, e-books made literary works more widely available and social networking made conversations about and ratings of literary authors and their works more accessible. Studies of the impact of literary authors might now have greater import than ever before.

[*] Corresponding author.
E-mail address: alonfriedman@usf.edu (A. Friedman).

[1] Created in 1967, WorldCat represents the catalog records of thousands of libraries. As of 2016, OCLC WorldCat had over 314 million bibliographic records, adding a record at an average rate of one every ten seconds (OCLC, 2015).

http://dx.doi.org/10.1016/j.lisr.2017.07.001
Received 16 December 2016; Accepted 4 July 2017
Available online 24 July 2017
0740-8188/ Published by Elsevier Inc.

One of the best-known bibliometric methods in the field of library and information science (LIS) is Lotka's law (Askew, 2008), which describes the frequency of publication by authors in any given field and has mainly been used to understand scientific writings rather than literature. The extension of this law to literature would be significant to the development of a bibliometric theory for the humanities and social sciences. This study explores the difference between scientific publication and popular literature as it pertains to the metrics of impact, and examines various recent attempts to rank literary authors according to different perspectives. To do so, this study focuses on the relevance of Lotka's law in examining the distribution of authorship in literature as it pertains to authors' impact. In particular, the study considers the following questions: (1) Is Lotka's law relevant to the world of literature? (2) What can a Lotkaian approach explain about the distribution of world literature? and (3) What bibliographical data ought to be collected and measured in examining literary rather than scientific eminence? This study will help librarians and those conducting research in LIS by examining evidence that Lotka's law can be used to measure the impact and productivity of literary authors.

3. Literature review

3.1. Lotka's law of scientific productivity

Research on author productivity has its origins in the work of the Austrian born American statistician Alfred J. Lotka (1880–1949). In 1926, Lotka investigated author publication productivity among physicists, using a decennial index of *Chemical Abstracts* and Aurbach's index to *Geschichtstafeln der Physik* (Aurbach, 1910; Lotka, 1926). Lotka determined that the volume of author production could be determined by counting the number of names in the index of *Chemical Abstracts* against the number of entries for each name. Lotka found that for each set of data, the points that represented the author's productivity were scattered closely around a strength line on a logarithmic scale. Lotka's law shows an asymmetric distribution with a concentration of articles among a few authors, while the remaining articles are distributed amid a larger amount of authors with low distribution. These findings had such profound implications about author productivity that they were later generalized as Lotka's law, one of a small number of bibliometric laws (Bookstein, 1976; De Bellis, 2009).

Lotka's law states that the number of authors making *n* contributions is about $1/\mathrm{n}^2$ of those producing single publications. The contributions of authors producing single publications comprise about 60% of the entire population in a specific field. Lotka's basic formula outlines the number of authors, represented as y_x, credited with x number of papers that appear inversely proportional to x, which is the output of each individual author. The relation is expressed as $\mathrm{X}^{\mathrm{n}}\ Y_x = C$ where y_x is the number of authors making x contributions to the subject and n and C are the two constants to be estimated for the specific set of data. Lotka noted that the equation applied to a variety of phenomena.

Lotka's law became a standard procedure in the field of information science when Pao (1985, 1986) established a testing and validation procedure to examine Lotka's law (Rai & Kumar, 2005). She outlined a testing procedure for Lotka's law that consisted of three steps: (1) data collection procedure, (2) estimation of the unknown parameters in the model, and (3) testing conformity of the observed data to the theoretical distribution by means of a goodness-of-fit test. Another important contribution made by Pao (1985) was the measurement of validity. Pao presented an evaluative framework for comparison of authorship data with Lotka's law's predictions to measure the validity of Lotka's law. This validation framework includes measurement of the variables and their tabulation, form of the model, and parameter estimation and criterion for goodness-of-fit. Pao recommended the Kolomgrov-Smirnov (K-S) as a form for evaluating the statistical significance of results. Appendix A summarizes Pao's six-step recommendations for applying Lotka's law.

However, a problem with Lotka's law, according to Askew (2008), is the lack of evidence of a clear and conclusive methodology supporting empirically validated data. Nicholls (1986, 1989) modified Pao's validation procedure for testing Lotka's law as a result. Despite this issue, the present study follows Pao's (1985) validation procedure, due to its popularity among researchers as a method of validating their study findings.

Another well-researched aspect of Lotka's law is the sample size of the data collection. Many studies using a small sample size found that their results did not conform to Lotka's law, leading Huber and Wagner-Dobler (2001) to recommend a larger sample size in order to reliably test Lotka's law. The breadth and scope of the source is also important. Typically, research studies testing Lotka's law have used $n = 2$ (Budd & Seavey, 1990; Murphy, 1973; Schorr, 1975) as the value of the exponent, which may have contributed to Lotka's law commonly being referred to as an inverse square law when calculating the value of C. While Lotka did present and discuss his formula in simpler terms using the value $n = 2$, it is important to note that he calculated the value of n (and C) for each set studied. Therefore, rather than referring to Lotka's law as the inverse square law, it would be more appropriate to refer to it as an inverse power law, since the value of n is calculated for each data set tested, and its value is not always equal to 2, as found in this study and a number of others (Egghe, 2005; Nicholls, 1989; Patra & Mishra, 2006; Rai & Kumar, 2005).

Lotka's law has also been criticized for not being able to support current academic research trends. According to Kretschmer and Rousseau (2001), in very large groups where researchers almost always collaborate with each other, each publication yields a credit to the same group of authors. This finding was supported by Tscharntke, Hochberg, Rand, Resh, and Krauss (2007), and many others, who reported that the increasing pattern of collaboration across scientific disciplines makes the issue of the sequence of contributors' names a major concern to academic evaluation committees in measuring their faculty's productivity.

3.2. Applications of Lotka beyond the sciences

Many academics and scientific researchers have employed Lotka's law to examine author productivity and publications. The potential of Lotka's law for application beyond the sciences led Egghe (2005) to coin the term “Lotkaian.” Of particular interest to Egghe was the explication of Lotka's exponent, α, in the formula $f(n) = C/n^{\alpha}$. The term Lotkaian captures the essence of the application in the present study of Lotka's law, substituting factors such as the number of works about an author for citations to the author, to analyze impact.

Murphy (1973) was the first to raise the question of whether Lotka's law could be applied to non-scientific productivity, although his own work only covered scientific journals. Bender (2008) took the next step by applying Lotka's law to museum catalogs. He reported that historical art catalogs were not suited to the study of the iconography of a specific subject across artists. He found that only special topical catalogs fit his study, while historical art catalogs were not optimally suited for studying the iconography of specific subjects across a range of artists.

The skewed distribution of publications found in science also applies to music, as can be seen by studying the artists who scored top-selling (gold and platinum) singles. Fox and Kochanowski (2004) analyzed the history of musical chart success with respect to the factors of musical grouping, gender, and ethnicity. They found that frequency distributions varied by race and gender, and that even where Lotka's law could not explain the empirical distribution, a generalized Lotkaian distribution provided a good model of music superstardom. This generalized distribution is $y_n/y_1 = 1/n^k$ where y_n is the number of artists, y_1 is the number of artists with one gold record, and k is a constant (Fox & Kochanowski, 2004, p. 516).

In Murray's (2003) examination of eminence in a broad range of endeavors, including literary writing, he took note of Lotka's law (though he does not consider it a "law"). Murray's approach does not use citation analysis, but instead follows the tradition of studies by early psychologists such as Galton (1869) and Cattell (1903) in measuring genius; Murray measures the amount of space allotted to figures in standard reference works. Following Woods (1911), Murray calls this approach historiometry. Murray devotes a chapter to the "Lotka curve," showing that great cultural achievement does not follow a normal distribution, which would look like a bell-shaped curve, but rather is concentrated at the top with a small number of individuals of extraordinary talent.

3.3. Differences between literary and scientific publications under the bibliometric paradigm

As defined by Glanzel and Schoepflin (1999), the term *bibliometrics* refers to the "application of mathematical and statistical methods to books and references" (p. 12). Such a definition suggests that bibliometric methods can and ought to be applied to any genre, subject matter, and vehicle of written communication. In practice, however, studies have focused almost exclusively on scientific communication in periodical literature. Indeed, the primary bibliometric methodology of counting citations of articles seems tailor-made for measuring the impact of scientific authors.

There is a growing body of literature on bibliometrics in the humanities (Kawamura, Thomas, Tsurumoto, Sasahara, & Kawaguchi, 2000; Nederhof, 2006) and many other disciplines, however, applying a bibliometric approach to a non-technical subject, such as literature, reveals certain problems in that approach. While literary and scientific texts share shelf space in libraries of various kinds, the two domains differ significantly in many respects. The cutting edge of science is found in articles (including many that are co-authored) in journals. In most cases, articles cite other earlier articles. The value of scientific literature can be understood partly through the output of the scientists who contribute to that literature, and partly through the citations of those papers by other scientists. Impact and influence, as well as the growth of research and the connections among researchers can be traced through citations.

Unlike scientific writings, which are aimed mainly at fellow professionals, the audience for literary writings consists of the public at large. They may read a work for pleasure or personal enlightenment or as part of their education (whether assigned, extracurricular, or self-directed), or they may not read the work at all, but rather see and hear the work in performance. Additionally, while the published journal article is universally accepted as the basic unit of communication in science, literary works exist in numerous genres, including novels, nonfiction, short stories, poems, criticism, essays (which may appear in magazines or specialized periodicals), speeches, plays, monologues or other performance pieces, songs, and more. This variety of formats, in terms of genre, publication, and delivery, raises the question of how to use ranking to evaluate literary works.

3.4. Library ranking for literary authors

Nowadays, relevance ranking has become a common method for presenting the results of a research query in library catalogs and on any web search engine. Those results are ranked algorithmically in terms of their relevance to the query, based on the search terms expressed in the document, and many other factors. According to Egghe and Rousseau (1990) and Garfield (1979), the growth of bibliographic data has received a boost from the revolutionary increase in computer power and the growing (now ubiquitous) production of information in digital form. This has led to the use of bibliographic data in a quantitative paradigm to measure the importance of journals, papers, programs, individual researchers, and disciplines. In ranking literary authors, Burt's (2001, 2009) and Bloom's (2002) rankings are used to analyze the data sets that were chosen based on their inclusion in recent books by Bloom (2002), Burt (2001, 2009), and Gottlieb, Gottlieb, Bowers, and Bowers (1998), all three of which attempt to rank authors in terms of their contribution to world literature and culture. Murray (2003) proposes a different score for the total accomplishment of many individuals, including authors. While Murray included thousands of authors from Arabic, Chinese, Indian, Japanese, and Western literature, his scores were used here only to corroborate the numbers provided by the other three surveys.

The major surveys intended to rank authors according to perceptions of their impact on the culture or literature, but all admitted to some subjectivity. Bloom's ranking is unabashedly personal, developed for the purpose of discussing his notion of genius. Bloom's idiosyncratic system of ten emanations (sephirot), each divided into two sets of five he calls lustres (a word he chose based on obscure literary connotations), derives from a combination of Kabbalah and Gnosticism, both of which have deeply influenced his thought (Baumlinn, 2000). Bloom's grouping is based on his own personal associations, and he insists on the very first page that the authors he selected for discussion are not "the top hundred," but only the ones he wanted to write about.

In the second revised edition of his book on the greatest authors, Burt explains his own approach and the skills and interests he brings to the project, writing,

Although I have taught the works of many of the writers in this ranking for > 25 years, I make no special claims to comprehensive expertise in the full range of world literature over the centuries. Rather, I have approached the task in the spirit of a general reader who is forced to choose, based on literary tradition, critical history, and personal preference, the best that has been written. I have, as best as I could, made choices that reflect some consensus beyond personal taste or a narrow cultural bias (Burt, 2009, p. xv).

A different approach than that of Bloom (2002) or Burt (2001, 2009) was presented by Gottlieb, Gottlieb, Bowers, and Bowers (1998). Their ranking of the most influential people of the second millennium C.E. does not focus on a single category of achievement such as literature, but is broadly based, including statesmen, generals, royalty, entrepreneurs, and tycoons as well as artists, scientists, inventors, and many others who have made a significant mark on human life, for good or ill.

4. Method

This study follows Pao's (1985) methodology, the original procedure employed by Lotka (1926), but with two modifications. The first is the modification of the sample size, and the second concerns data collection under the category of "fame". Under the sample size outlined according to Lotka's law, data collection is intended to demonstrate population distribution, in order to identify where production was concentrated. While the foundation of Lotka's law is concerned with measurement of author productivity in the sciences, the literary universe is aimed mainly at a general and professional audience. As a result, it was a challenge to capture the entire universe of literature. In the absence of a list of all literary authors whose works are found in WorldCat, this study could not use simple random sampling. Therefore, a non-probability method of convenience sampling was used to capture data. This sample technique based on the judgment of the researcher can be used when the entire data set cannot be accessed, according to Lavrakas (2008).

The data in this study was collected at two different times. The first stage was conducted in 2007, the second in August 2014. Data were collected on the number of works by literary authors, as evidenced by the 100 (main entry, personal name) or 700 (added entry, personal name) fields of the MARC record in OCLC, and on the number of works about those authors, as evidenced by the MARC 600 (subject added entry, personal name) field. Data were also collected on the number of works both by and about an author, those by but not about an author, and those about but not by the author, using Boolean search principles as outlined by Naun (2010). In order to collect the data, a data collection technique called "fame" was employed. Martindale (1995) employs this technique in analyzing literature's impact by counting the works devoted to a given author. In the present study the impact and eminence of literary authors was measured by examining the number of bibliographic records found in OCLC WorldCat linked to the names of eminent authors during the two different stages of data collection in 2007 and 2014.

In the text that follows, readers should keep in mind that the data collection was based on convenience sampling and does not represent or capture the entire population of literary authors found in the OCLC WorldCat catalog. The focus of the study was to assess Burt's, Bloom's, and Lotka's frameworks by examining the data from 2007 and 2014. This approach does not allow for generalization about the population. In order to validate the study's scores, the authors used a K-S test at the level of significant of 0.10.

5. Results

In the first analysis, the study employed Burt's and Bloom's ranking. The authors have broken down their formula into five factors: lasting influence (41.7%), effect on the sum total of wisdom (c. 20.3%), influence on contemporaries (c. 16.7%), singularity of contribution (12.5%), and charisma (c. 8.3%). It is possible for different judges to disagree on these factors and difficult to know how to give them all numerical scores. Although the particular factors chosen for scoring seem reasonable, no justification is given for the specific ratios, which seem arbitrary and odd, and leads to questions about why such specific ratios were chosen. In sum, Burt's and Blooms's approach, though it uses numbers, cannot really be called quantitative since at heart it relies on combined hunches. Despite this apparent shortcoming of their approach, the present study finds that their ranking comes closest to matching measurements based on OCLC data.

This study precludes a Lotkaian framework for reading Burt's and Bloom's rankings, due to a concern that Burt and Bloom did not provide clear definitions and parameters to convert their models to numerical analysis. Table 1 represents a sample of 11 authors, of the 1000 for which data about author productivity was collected in OCLC WorldCat. In addition, to collect the WorldCat records, the data were sorted by Gottlieb et al.'s (1998) ranking.

Due to the creation of a large number of records for editions of existing works in e-book formats in the years between 2007 and 2014, the numbers were significantly higher when the same authors sampled seven years previously were checked again in 2014. The greatest difference occurs in publications about, rather than by, an author. In 2014, WorldCat displayed 785 more references than in 2007. It is interesting to note that under Gottlieb et al.'s ranking framework, measures remain the same during both years (2007 and 2014). Fig. 1 represents the difference between the results from 2007 and 2014.

Table 1
"By," "about," and "by and about" data from a convenience sample of authors in 2007 and 2014.

Year	Gotlieb et al.	Author	By	About	By and About
2007	15	Dickinson, Emily	103	7665	446
2014	15		119	7761	654
2007	19	Ibsen, Henrik	203	6521	734
2014	19		304	15,321	832
2007	30	Dante Alighieri	237	17,312	1395
2014	30		273	13,212	1375
2007	34	Tolstoy, Leo	282	5932	652
2014	34		285	5943	544
2007	36	Voltaire	297	4946	686
2014	36		321	5212	701
2007	44	Joyce, James	364	6360	517
2014	44		342	6359	527
2007	53	Milton, John	431	7834	730
2014	53		474	8012	763
2007	62	Hawthorne, Nathaniel	470	6677	540
2014	62		427	6856	551
2007	70	Dickens, Charles	506	9378	1259
2014	70		579	10,121	1369
2007	131	Twain, Mark	786	21,017	3529
2014	131		784	21,186	3631

Lotka's law is calculated as $X^n \ Y_x = \mathrm{C}$, where Y_x is based on the number of authors, each credited with x number of manuscripts, and is inversely proportional to X, which is the output of each individual author. The two constant values in Lotka's law, n and C, stand for estimates for the specific set of data. Lotka's original 1926 studies found that the values of n were 2.02 for *Chemical Abstracts* data and 1.888 for the *Geschichstafen der Physik* data. The present study calculates the value of n by using the least square-method to estimate the best value for the slope of a regression line that is the exponent n for Lotka's law. The slope is usually calculated without data points representing authors of high productivity. Since the values of the slope change with different numbers of points for the same set of data, the value of $n = 2$ is used, which will be identified as the best slope for the observed distribution. The analysis results in a value of n as -1.420903 for 2007 and -1.2543 for 2014.

Due to the above results, the non-negative fractional values of n were employed, and the summation of the series can be approximated using a function that calculates the sum of the first P term. Using the value of n, the next step was to calculate the value of C. For 2007, the constant C was equal to 0.6908, in comparison to 2014, when the value of C was 0.976. These findings allow the calculation of exponent n without pairing the data. Table 2 captures the calculation of exponent n during 2007 v. 2014.

The next stage of the analysis was calculating the Kolmogorov–Smirnov (K-S) goodness-of-fit test, as recommended by Pao (1985), to ensure the results were accurate. A K-S analysis, conducted to compare the distributions of the observed and expected values of y for the literature, indicated no significant difference in the two distributions ($p < 0.000$). The difference between the two distributions was 1.43 with a mean of 0.86.

Next, the value of D max was calculated. The critical value of 0.01 at the level of significance was calculated. The result for 2007 was equal to 0.1317786, whereas the result for 2014 was 0.2288. No significant differences from the theorized distribution were found in either case. The maximum deviation for 2007 equaled 0.13177, which exceeds the critical value of 0.13177 at the 0.01 level of significance. For the second data set, from 2014, the maximum deviation equaled 0.228, which also exceeded the critical value of 0.01 level of significance. Therefore, both distributions fit into Lotka's law. Fig. 2 captures the two distributions.

Data on Gottlieb et al.'s (1998) rankings do not fit neatly with predicted Lotkaian distributions. The analysis by Gottlieb et al. reveals no major differences between 2007 and 2014. Next, calculating the values of constant C and exponent n shows that the value of constant C is 0.9342 and the value of n exponent is 1.4454.

The validity of Lotka's law as a methodology has been discussed by many researchers. Sen, Taib, and Hassan (1996) conclude that Lotka's law is applicable to the field of library and information science, measuring the annual index of Library and Information Science Abstracts (LISA) as a test case. In the current study, the Kolmogorov–Smirnov (K-S) goodness-of-fit test was also conducted to determine if Lotka's law can be used as a reliable tool to predict literary author publication productivity from the observed values. Conover (1971) notes the K-S test is more powerful than the 2-test, and is an appropriate test for ranked data.

Specifically, this study conducted K-S analysis, at the 0.10 level of significance, to compare the distributions of the observed and expected values of y for the literary authors. The test indicated no significant difference in the two distributions ($p < 0.000$). The K-S test uses the maximum vertical deviation between the two curves as the statistic D max. The values of D max with regard to 2007 data, 2014 data, and

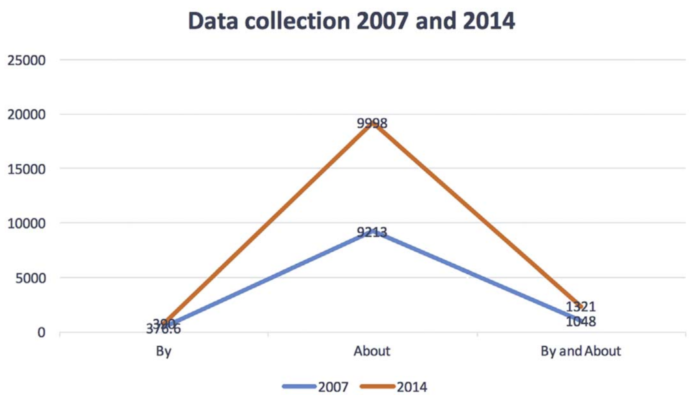

Fig. 1. Data collection summaries: 2007 v. 2014.

Gottlieb et al. were equal to 0.1317786, 0.2288, and 1.1 respectively. Tables 3 and 4 represents the results from 2007 and 2014 in table formats. Fig. 3 compares the three distributions of 2007 data, 2014 data, and Gottlieb et al. (1998). The graph line in red represents the Gottlieb et al. (1998) distribution. As seen in Fig. 3, their ranking does not match well with the findings using OCLC data from 2007 or 2014. The value of constant C and exponent n with the D max value reveals that Gottlieb et al.'s (1998) ranking does not provide a good fit for author impact.

To deepen the focus of the study, the authors followed Pao–Lotka procedures that led to the following findings: Gottlieb et al.'s theory did not provide good predictions with reliable results of author literary

Table 2
The calculation of exponent n during 2007 and 2014.

Recording time	Author name	Books	Log x	Log y	x_y	x^2
2007	Balzac, Honoré de	1000	3	1.4771213	4.43124	09
2014		1000	3.000321	1.6654321	4.76543	09.21
2007	Baudelaire, Charles	1000	3.000321	1.6654321	4.76544	09.23
2014		1000	3.000321	1.6654321	4.76543	09.23
2007	Cervantes, Saavedra, Miguel de	1000	3.000323	1.6654321	4.76543	09.23
2014		1000	3.00032	1.6654321	4.76543	09.23
2007	García Lorca, Federico	2000	3.3012	1.6627578	5.48813	10.8968
2014		2000	3.43012	3.43012	5.54321	11.23355
2007	Mann, Thomas	2000	3.45013	1.885432	5.65433	11.5687
2014		2000	3.2154	1.766732	5.05434	10.8765
2007	Faulkner, William	2000	3.21567	1.77721	5.05677	10.1234
2014		2000	3.23457	1.7999	5.07753	10.2124
2007	Wilde, Oscar	2000	3.45633	1.8976	5.03221	10.3214
2014		2000	3.56789	1.9876	5.04567	10.3567
2007	Eliot, George	3000	3.67898	1.39794	4.86007	12.09037
2014		3000	3.78902	1.6543209		
2007	Chekhov, Anton Pavlovich	4000	3.60206	1.413638	5.15585	12.97484
2014		4000	3.66432	1.432156	5.32124	13.01246
2007	Byron, George	5000	3.69897	1.30103	4.81242	13.68238
2014		5000	4.00023	1.790453	5.00012	13.54328
2007	Becket, Samuel	6000	3.77815	1	3.77815	14.27443
2014		6000	3.89873	1.032	4.21236	14.65432
2007	Molière	7000	3.84509	1.041397	4.00425	14.27443
2014		7000	3.96642	1.032274	4.00321	14.11654
2007	Tolstoy, Leo Graf	9000	3.95424	0.69897	2.76389	15.63603
2014			4.01236	0.989832	3.12235	16.03257
2007	Joyce, James	10,000	4	0.778151	3.1126	16
2014			3.87765	0.654321	2.89765	15.00322
2007	Frost, Robert	11,000	4.04133	0.301021	1.216458	16.33365
2014			5.23579	0.204543	1.65543	16.99956
2007	Woolf, Virginia	12,000	4.07918	0.477213	1.94626	16.63972
2014			4.06901	0.466321	1.87543	16.56789
2007	Austen, Jane	13,000	4.14613	0.3001	1.23842	16.92452
2014			4.65428	0.212345	1.12234	17.00232
2007	Hugo, Victor	14,000	4.14613	0	0	17.19038
2014			4.15443	0.004322	0.00689	17.24325
2007	Kafka, Franz	16,000	4.20412	0	0	17.67462
2014			4.23568	0	0	17.68546
2007	Williams, Tennessee	20,000	4.30103	0	0	18.4986
2014			4.45	0	0	18.8642
2007	Christie, Agatha	25,000	4.39794	0	0	19.34188
2014			4.45007	0	0	21.04334
2007	Hemingway, Ernest	64,000	4.46021	0	0	23.09936
2014			4.23178	0	0	23.09921
2007	Shakespeare, William	64,000	4.80618	0	0	23.09937
2014			4.86043	0	0	23.98643

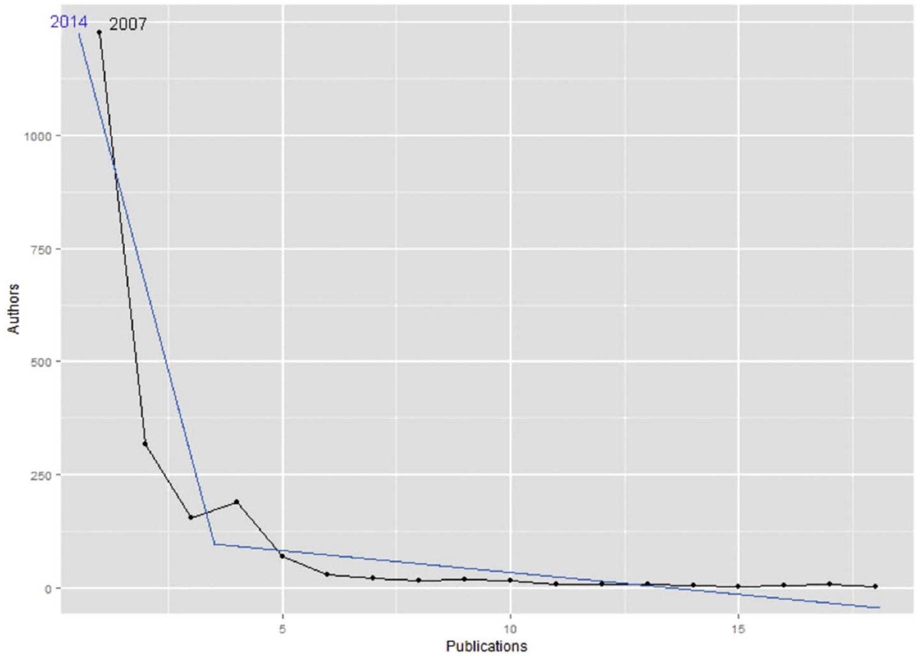

Fig. 2. Lotka' s law Distribution based on 2007 v. 2014.

Table 3
The Kolmogorov-Smirnov (K-S) goodness-of-fit test results for 2007.

Row number	Authors	% Authors	Cum Sum of % Authors	Expected % Authors	Cum Sum of expected % of authors	D
1	50	0.326797386	0.326797386	0.064820973	0.064820973	0.261976412
2	10	0.065359477	0.392156863	0.048065064	0.112886037	0.279270826
3	25	0.163398693	0.555555556	0.043382124	0.156268161	0.399287394
4	9	0.058823529	0.614379085	0.023914054	0.180182216	0.434196869
5	12	0.078431373	0.692810458	0.021181798	0.201364013	0.491446444
6	5	0.032679739	0.725490196	0.018496817	0.21986083	0.505629366
33	8	0.052287582	0.777777778	0.015313563	0.235174393	0.542603385
62	11	0.071895425	0.849673203	0.014330487	0.24950488	0.600168323
88	7	0.045751634	0.895424837	0.014091716	0.263596596	0.63182824
100	6	0.039215689	0.934940523	0.00887888	0.272528484	0.663212038

Table 4
The Kolmogorov-Smirnov (K-S) goodness-of-fit test results for 2014.

Row number	Authors	% Authors	Cum Sum of % Authors	Expected % Authors	Cum Sum of expected % of Authors	D
1	50	0.42011111	0.4190046	0.0803245	0.069032	0.2836789
2	10	0.085457466	0.4676542	0.0670432	0.142666	0.6346563
3	25	0.24454321	0.8545349	0.0689724	0.196432	0.6543429
4	9	0.08653256	0.85432346	0.0659842	0.2654228	0.6786423
5	12	0.078431373	0.692810458	0.0456717	0.4356783	0.5613467
6	5	0.08325799	0.87563422	0.3245324	0.2078997	0.678
33	8	0.0645676	0.3456798	0.0543207	0.659064	0.6890453
62	11	0.06595342	0.6789065	0.0234578	0.0467903	0.706543
88	7	0.06596534	0.990321	0.060543238	0.4789055	0.7890432
100	6	0.06543256	0.87609676	0.016547903	0.35789877	0.87645328

publication productivity using OCLC. The study finds that Lotka's law can be used to measure literary author publication productivity with reliable results. It also conducted a K-S goodness-of-fit test to measure the validation of Lotka's law.

6. Discussion

6.1. Literary authors and bibliographic impact

This study contributes to an understanding of the relative fame or bibliographic impact of literary authors. It used a bibliometric approach devised for studying the impact of scientific authors, but adapted for the purpose of studying literary authors and their works, since literature makes its impact on culture and the larger reading public in a manner quite differently than science. While the influence of science can be seen through citations of articles by other scientists, literature achieves its impact through analysis, literary biography, reproduction in new editions, and recreation and performance in new formats. Therefore, instead of focusing on articles in professional journals and citations of them, this study counts bibliographic records of whole works cataloged

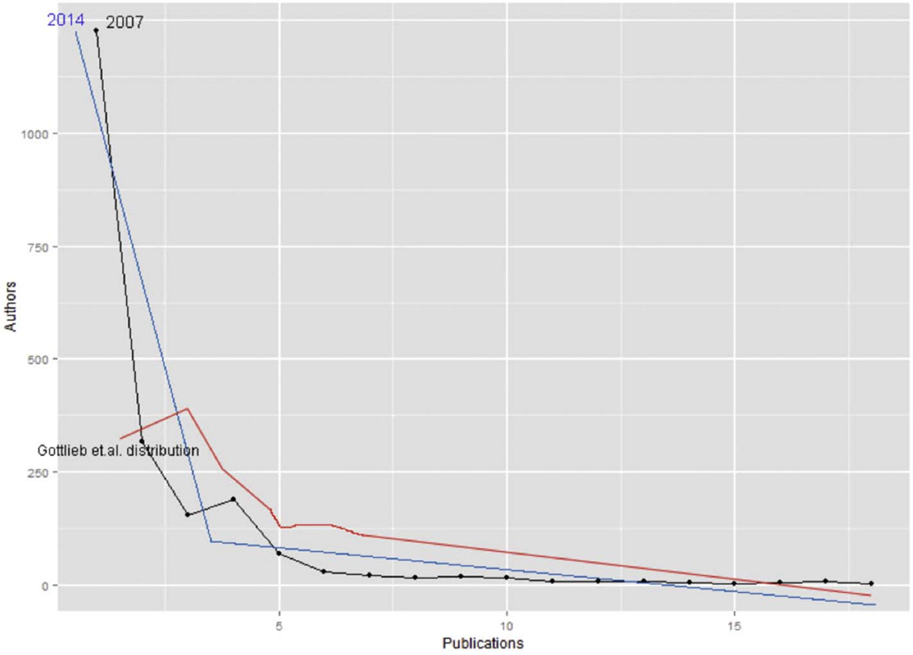

Fig. 3. Comparison 2014, 2007 and Gottlieb et al. distributions.

as created by authors, as well as works cataloged about those authors; this approach prioritizes book-length works over articles in periodicals. Such a focus may be profitably employed in a broad range of book-based disciplines in the humanities and social sciences.

The analysis of distributions for works both by and about these major authors conforms to a Lotkaian interpretation. This model enables calculation for values of the C constant equal to 0.6908 for 2007 and 0.976 for 2014, with exponent n equaling 1.420903 for 2007 and -1.2543 for 2014.

Beyond the findings strictly about the applicability of Lotka's law to literature, the potential value of bibliographic records rather than citations in bibliometric studies was demonstrated. Such materials, it may be argued, are more pertinent to literary fame or impact than the number of journal articles and citations of those articles. In particular, the study considered the number of records for works by, about, by and about, and by but not about the authors. Using the adjustments for C and n discussed, it confirms the validity of a Lotkaian pattern applicable to the major literary authors. This finding helps build an understanding of the structure of the domain of world literature within the larger universe of cultural productions.

6.2. Limitations

Unlike scientific data, where the common scope of the bibliographic data sets often allow researchers to look at a list of a single journal or multiple journals, this study employs the OCLC WorldCat catalog by drawing on the data of literary author rankings based on Bloom (2002), Burt (2001, 2009) and Gottlieb et al. (1998). Data were collected during two different periods of 2007 and 2014. Literary authors, unlike scientific authors, do not have a single source of measurement and as a result, ranking was used as the methodology. Due to the nature of the source, no generalization about the entire population of literary authors found in OCLC WorldCat was possible. Also, it must be noted that this study does not take into account the journal literature, which would be likely to include many articles about these authors, though not by them.

7. Conclusion

Since Lotka's discovery 90 years ago of power laws pertaining to the relative productivity of scientists, most researchers who have followed up on his work, developing the burgeoning field of bibliometrics since the 1950s, have concentrated on technical and academic publications in an environment that has increasingly shifted to multi-author collaboration. This study demonstrates the applicability of the same laws to publications by non-scientific authors with a general readership. It demonstrates the value of a method based on using OCLC data on records by and about authors, combined with a Lotkaian approach to determine authorial impact. This research can apply to a much wider spectrum of literature in collections characterized by power laws.

The pertinence of such research to library and information science is apparent not only because libraries of many kinds maintain the bulk of resources in and about literature, but because the public still relies on libraries (academic and public) for access to these materials. Notions about literary canons are important in collection management, with practical applications for sorting literature and authors. A study such as this, using quantitative data, can verify the adequacy of subjective rankings and qualitative studies of author merit and cultural consecration.

Patterns can be observed from changes over time in the set of records for works by and about authors. The two moments in time captured by the study are characterized by developments in bibliographical technology, most notably the popularization of electronic books, contributing to changes found in the patterns of author impact. Focusing on changes over time enables librarians to determine whether technology has improved access to literature and how libraries can improve their services to meet the needs of patrons.

With more and more digital production and reproduction of literary works, as well as more reading occurring online, it remains to be seen whether Lotka's law will continue to apply to the new and evolving ways of measuring and reading online, including reading habits in different genres of writing, including literature. Future studies will need to address the possible application of this law to the metrics of blogs, Twitter, and new ways of disseminating and consuming literature that have yet to be invented.

Appendix A

Pao's six-steps procedure for applying Lotka's law.

1. Measurement and tabulation: the number of authors' y_x contributing x paper is organized into a size frequency table of n, x, y pairs.
2. Model: the generalized inverse-power model where, $y_x = kx^{-b}$ is adopted.
3. Estimation of slope b: The ordinary linear least squares estimate of b in the transformed model:

$\log y_x = \log$ K–b $\log_x$, x = 1, 2, x_{max}

4. Estimation of constant C:

Based on $y_x = c/x^n$

Pao (1985) recommend dividing both sides of equation by $\Sigma y_{x,}$ the total number of authors

$y_x/\Sigma y_x = (c/\Sigma y_x)(1/X^n)$

Let $f(y_x) = y_x/\Sigma y_x$ provides the fraction of authors making x contributions and $C = c/\Sigma y_x$ is the new constant, expressed as a fraction of the total sample of authors. Thus, equation $y_x/\Sigma y_x = (c/\Sigma y_x)(1/X^n)$ can be written as

$f(y_x) = C(1/x^n)$

According to Pao (1985), this equation is another form of Lotka's general law that stands for the percentage of authors $f(y_x)$, where each with x is the number of publications. This is inversely proposal to x raised to the nth power.

5. Extrapolating from Lotka's calculation of the special case for $n = 2$, the general formulation equation for any value of n is as follows

$y_1 = c(1/1^n)$

$y_2 = c(1/2^n)$

$y_3 = c(1/3^n)$

$y = c(1/X^n)$

Summing both sides of these equations will provide us the following formula where, according to Pao (1985), we need to divide both sides by the total number of authors

$\Sigma y_x = c(1/1^n + 1/2^n + 1/3^n + 1/X^n)$

$\Sigma y_x/\Sigma y_x = (c/\Sigma y_x)(\Sigma 1/x^n)$

Since the summation of “Σ” and yx together with c/“Σ” $y^x = C$ allow us to generate the following equation: C = (1/“Σ” *$1/x^n$), according to Pao (1985, 121–134) and Nicholls (1989).

6. Test: There are several statistical tests available for goodness-of-fit. Among those tests, is the Kolmogorov–Smirnov (K-S) test. The aim of this test is findings the theoretical cumulative frequency distribution by comparing it with the observed cumulative frequency distribution. The point at which the two observed distributions show the maximum deviation can be determined. The point at which the two observed distributions show the maximum deviation can be determined. The null hypothesis is then rejected if the calculated value of D is greater than critical value, according to Corder and Foreman (2014).

References

Askew, C. (2008). *An examination of Lotka's law in the field of library and information studies*. (Doctoral dissertation). Available from ProQuest Dissertations and Theses. (UMI No. 3388194).

Aurbach, F. (1910). *Geschichtstafeln der Physik [History of physics]*. Leipzig, Germany: Barth.

Baumlinn, J. S. (2000). Reading bloom: Or, lessons concerning the “reformation” of the western literary canon. *College Literature*, *27*(3), 22–46.

Bender, K. (2008). *Lotka's law of productivity*. Retrieved from https://sites.google.com/site/venusiconography/home/research-papers/lotka-s-law-of-productivity.

Bloom, H. (2002). *Genius: A mosaic of one hundred exemplary creative minds*. New York, NY: Warner Books.

Bookstein, A. (1976). The bibliometric distributions. *Library Quarterly*, *46*(4), 416–423.

Budd, J. M., & Seavey, C. A. (1990). Characteristics of journal authorship by academic librarians. *College & Research Libraries*, *51*(5), 463–470.

Burt, D. S. (2001). *The literary 100: A ranking of the most influential novelists, playwrights, and poets of all time*. New York, NY: Facts on File.

Burt, D. S. (2009). *The literary 100: A ranking of the most influential novelists, playwrights, and poets of all time* (Rev. ed.). New York, NY: Checkmark Books.

Cattell, J. M. (1903). A statistical study of eminent men. *Popular Science Monthly*, *62*, 359–376. Retrieved from https://en.wikisource.org/wiki/Popular_Science_Monthly/Volume_62/February_1903/A_Statistical_Study_of_Eminent_Men.

Conover, W. J. (1971). *Practical nonparametric statistics*. New York, NY: Wiley & Sons.

Corder, G. W., & Foreman, D. I. (2014). *Nonparametric statistics: A step-by-step approach*. NY: John Wiley & Sons.

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the Science Citation Index to cybermetrics*. Lanham, MD: The Scarecrow Press, Inc.

Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. Amsterdam, Netherlands: Elsevier.

Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Amsterdam, Netherlands: Elsevier.

Fox, M. A., & Kochanowski, P. (2004). Models of superstardom: An application of the Lotka and Yule distributions. *Popular Music & Society*, *27*(4), 507–522.

Galton, F. (1869). *Hereditary genius: An inquiry into its laws and consequences*. London, England: Macmillan.

Garfield, E. (1979). A statistically valid definition of bias is needed to determine whether the Science Citation Index discriminates against third world journals. *Current Science*, *73*(8).

Glanzel, W., & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information Processing and Management*, *35*(1), 31–44.

Gottlieb, A. H., Gottlieb, H., Bowers, B., & Bowers, B. (1998). *1,000 years, 1,000 people: Ranking the men and women who shaped the millennium*. New York, NY: Kodansha America.

Huber, J. C., & Wagner-Dobler, R. (2001). Scientific production: A statistical analysis of authors in mathematical logic. *Scientometrics*, *50*(2), 323–337.

Kawamura, M., Thomas, C. D. L., Tsurumoto, A., Sasahara, H., & Kawaguchi, Y. (2000). Lotka's law and productivity index of authors in a scientific journal. *Journal of Oral Science*, *42*(2), 75–78.

Kretschmer, H., & Rousseau, R. (2001). Author inflation leads to a breakdown of Lotka's law. *Journal of the American Society for Information Science and Technology*, *52*, 610–614.

Lavrakas, P. J. (2008). *Encyclopedia of survey research methods*. Thousand Oaks, CA: Sage.

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

Martindale, C. (1995). Fame more fickle than fortune: On the distribution of literary eminence. *Poetics*, *23*, 219–234.

Murphy, L. J. (1973). Lotka's law in the humanities? *Journal of the American Society for Information Science and Technology*, *24*, 461–462.

Murray, C. (2003). *Human accomplishment: The pursuit of excellence in the arts and sciences, 800 B.C. to 1950*. New York, NY: HarperCollins.

Naun, C. C. (2010). Next generation OPACs: A cataloging viewpoint. *Journal Cataloging & Classification Quarterly*, *48*(4).

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, *66*, 81–100.

Nicholls, P. T. (1986). Empirical validation of Lotka's law. *Information Processing & Management*, *22*, 417–419.

Nicholls, P. T. (1989). Bibliometric modeling processes and the empirical validity of Lotka's law. *Journal of the American Society for Information Science*, *40*, 379–385.

OCLC. (2015). *A global library resource*. Retrieved from http://www.oclc.org/worldcat/catalog.en.html.

Pao, M. L. (1985). Lotka's law: A testing procedure. *Information Processing & Management*, *21*, 305–320.

Pao, M. L. (1986). An empirical examination of Lotka's law. *Journal of the American Society for Information Science*, *37*, 33–36.

Patra, S. K., & Mishra, S. (2006). Bibliometric study of bioinformatics literature. *Scientometrics*, *67*, 477–489.

Rai, L. P., & Kumar, N. (2005). A rationale for Lotka's law: An examination of empirical data. *Annals of Library & Information Science*, *52*(3), 103–107.

Schorr, A. E. (1975). Lotka's law and map librarianship. *Journal of the American Society for Information Science*, *26*, 189–190.

Sen, B. K., Taib, C. A., & Hassan, M. F. (1996). Library and information science literature and Lotka's law. *Malaysian Journal of Library & Information Science*, *1*(2), 89–93.

Tscharntke, T., Hochberg, M. E., Rand, T. A., Resh, V. H., & Krauss, J. (2007). Author sequence and credit for contributions in multi-authored publications. *PLoS Biology*, *5*(1), 13–14.

Woods, F. A. (1911). Historiometry as an exact science. *Science*, *33*, 568–574.

Alon Friedman is an assistant professor at the School of Information, University of South Florida. He received his PhD from the Palmer School, Long Island University, New York. His research interests and expertise focus on classification and visualization. He has worked as a web programmer for high tech startup companies. Dr. Friedman has published in *Journal of Documentation* and *Knowledge Organization*, among others, and is currently working on a new textbook on statistics and visualization using open source R.

Jay H. Bernstein died on July 11, 2016, in Brooklyn, New York. He became a librarian following a career as an anthropologist that included three years of ethnographic fieldwork in Borneo. He received his master's and doctoral degrees in anthropology from the University of California, Berkley, and an MLS from St. John's University, New York. From 2004, Dr. Bernstein worked as a librarian faculty member at Kingsborough Community College, Brooklyn, New York. He was an active member of the International Organization for Knowledge Organization, and published in *Advances in the Study of Information and Religion*, *American Anthropologist*, *Journal of Ethnobiology*, *Journal of Research Practice*, *Knowledge Organization*, and *Library Resources & Technical Services*.