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

A B S T R A C T

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 $1/n^2$ 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 $X^n Y_x = C$ 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