Empirical Laws, Theory Construction and Bibliometrics

DANIEL O. O'CONNOR
HENRY VOOS

BIBLIOMETRICS HAS COMMANDED the attention of numerous individuals in library and information science. The measurement of bibliographic information offers the promise of providing a theory that will resolve many practical problems. It is claimed that patterns of author productivity, literature growth rates and related statistical distributions can be used to evaluate authors, assess disciplines and manage collections. Yet, it is unclear if bibliometrics is merely a method or if it meets the test of a theory in its ability to explain and predict phenomena. This paper examines the properties of bibliometric distributions in a nontechnical manner.

Twelve years ago, Pritchard coined the term *bibliometrics* and defined it as "the application of mathematics and statistical methods to books and other media of communication."[1] Its purpose was:

> 1. To shed light on the processes of written communication and of the nature and course of development of a discipline (in so far as this is displayed through written communication), by means of counting and analyzing the various facets of written communication...;
> 2. The assembling and interpretation of statistics relating to books and periodicals...to demonstrate historical movements, to determine the national or universal research of books and journals, and to ascertain in many local situations the general use of books and journals.[2]

Daniel O. O'Connor is Assistant Professor, and Henry Voos is Professor, Graduate School of Library and Information Studies, Rutgers University, New Brunswick, New Jersey.

Both of these purposes emphasize that bibliometrics is primarily a method. The scope of bibliometrics includes studying the relationship within a literature (e.g., citation studies) or describing a literature.[3] Typically, these descriptions focus on consistent patterns involving authors, monographs, journals, or subject/language. The literature of bibliometrics is growing rapidly and a recent bibliography lists 2032 entries,[4] while another announced bibliography has 600 entries covering the years 1874 through 1959.[5]

Two concerns have occupied much of the bibliometric literature: an emphasis on mathematical or statistical methods, and a search for theoretical propositions. Fairthorne, Price and Bookstein have stated that there is great consistency among the various bibliometric distributions.[6] The Bradford, Lotka and Zipf distributions are considered the basic laws of bibliometrics,[7] and each of these distributions was empirically derived. The distributions are similar to each other as special cases of a hyperbolic distribution. Fairthorne summarized the similarities of the bibliometric distributions in 1969: "Almost all of them, whatever their starting-point, end with some kind of hyperbolic distribution in which the product of fixed powers of the variables is constant. In its simplest discrete manifestation an input increasing geometrically produces a yield increasing arithmetically."[8]

Thus, the similarities of the Lotka, Bradford and Zipf distributions are not surprising. These distributions are based on rank-order frequencies (or rank-size relations) where objects are classified and then ranked. Zipf found that rank times size equals a constant. As derived in a more general form by Mandelbrot, frequency of occurrence is a function of constants applied to size and rank.[9] Similar distributions emerge in describing the following phenomena: rivers, populations of cities, biological genera, books (ranked by number of pages), author productivity, citations to journals, and frequency of words.[10]

Relationship Between Empirical Laws and Theories

The occurrence of dissimilar events at constant rates may allow for prediction of the frequency of events, but it does not explain their causes.[11] There is no reason to assume that the ability to make empirical predictions will eventually lead to theoretical explanations. This philosophical issue has been dealt with by Carnap:

> ...theoretical laws cannot be arrived at simply by taking the empirical laws, then generalizing a few steps further. How does a physicist arrive at an empirical law? He observes certain events in nature. He

> notices a certain regularity. He describes this regularity by making an inductive generalization. It might be supposed that he could now put together a group of empirical laws, observe some sort of pattern, make a wider inductive generalization, and arrive at a theoretical law. Such is not the case.[12]

Carnap further states that generalization from observations will never produce a theory; instead, a theory arises "not as a generalization of facts but as a hypothesis."[13] Fairthorne addressed this problem in bibliometrics: "I have surveyed the hyperbolic laws as a whole, with bibliometric applications as particular cases. This unifies the formal aspects of this type of behavior, and collects tools for dealing with it, without invoking any hypothesis about the proximate causes of such behavior."[14]

Price has proposed a general bibliometric theory based on a hyperbolic curve, which he has named the Cumulative Advantage Distribution.[15] In speculating on the reasons for this distribution, Price makes a valuable contribution to concept formation and theory construction in bibliometrics. However, his Cumulative Advantage Distribution would be subject to Rapoport's criticism of similar rank-size laws:

> Clearly, if objects can be arranged according to size, beginning with the largest, *some* monotonically decreasing curve will describe the data. The fact that many of these curves are fairly well approximated by hyperbolas proves nothing, since an infinitely large number of curves resemble hyperbolas sufficiently closely to be identified as hyperbolas. No theoretical conclusion can be drawn from the fact that many J curves look alike. Theoretical conclusions can be drawn only if a rationale can be proposed that implies that the curves *must* belong to a certain class. The content of the rationales becomes, then, the content-bound theory.[16]

As Rapoport later points out, it is the classificatory procedure that is important along with the prior expectations of the classifier.[17] Hill identifies three sources of uncertainty in such statistical laws: "First, the probabilistic mechanism by which the population frequencies...are determined; secondly, the method of sampling from the population; thirdly, the way in which the sample is classified."[18] Thus, it is doubtful that the similarities of the various bibliometric distributions have great theoretical importance.

None of this denies the practical utility of applying bibliometric distributions to library problems, but it does bring into question two concerns: (1) the generality of bibliometric techniques, and (2) the likelihood that the bibliometric patterns will change over time. Although Line has denied many of the practical claims attributed to bibliometrics,[19] Broadus has applied citation analyses to collection building.[20] Other applications to collection management can be found in a special bibliometrics issue of *Collection Management* edited by Moll.[21] The widespread application of practical bibliometric methods—useful to library managers—will continue to be limited until a more general, unified theory is developed. Such a theory should allow for the possibility of change in bibliometric distributions. Hill stated that: "Zipf's law for city sizes has held until very recently, but the development of suburbia seems to have altered matters to a certain extent. A more sophisticated model...would deal with the dynamics of the situation, and not merely the one-dimensional view obtained at a given point in time."[22] A similar limitation could apply to the long-term stability of bibliometric distributions, and this might account for the minor differences in the distributions associated with various disciplines.

Another limitation of bibliometric distributions is the use of unidimensional descriptions of consistency in author productivity or journal citation patterns. The more popular, library-related areas of bibliometrics—Lotka and Bradford—are based on plotting one or two variables which are then reduced to a single dimension. Such descriptive analyses usually lack explanatory power, since there are not enough variables to posit that one event causally influences the outcome of another event. If bibliometric distributions have identifiable causes, then multidimensional analyses may provide more fruitful avenues of research than plotting new hyperbolic distributions. This multidimensional issue has serious implications for the sustained relevance of bibliometric distributions as aids to library decision-making. This does not deny the immediate usefulness of some of these distributions, but it does bring into question their explanatory power and their ability to generate new theoretical hypotheses. Two of these distributions—Lotka and Bradford—will now be examined in more detail.

The Lotka and Bradford Distributions

The Lotka distribution is based on an inverse square law where the number of authors writing n papers is $1/n^2$ of the number of authors writing one paper. Each subject area can have associated with it an exponent representing its specific rate of author productivity.[23] But this does not explain why one individual produces dozens of published papers on a subject, another individual produces several papers, and a third individual produces none. The variability of author productivity could be partly explained by each individual's background (e.g., schools attended, influence of mentors), current information environment (e.g., access to current publications, colleagues, libraries), and other characteristics.[24] The individual's affiliation with a particular discipline could establish different expectation levels for author productivity. For example, it is estimated that scientists produce an average of 3.8 articles per year, while those in the social sciences produce only an average of 0.5 articles per year.[25]

It could be proposed that author productivity is a function of many causes, and these might be grouped into two major conceptual areas: (1) an author's personal characteristics (e.g., intelligence, achievement, personality, expectations); and (2) the author's environment or situation (e.g., colleagues, availability of information, the problem under investigation, author's field or discipline). In addition, the interactions among personal characteristics and environmental characteristics would create a third conceptual area for future study.[26] Numerous variables could be developed from these three conceptual areas while recognizing that the point of this is to recast author productivity as something that is more than a univariate statistical distribution. Author productivity can be viewed as having a multitude of preconditions which cause authors to behave in different ways. It is assumed that the variability in these causes is systematically related to the variability in productivity. In the building of causal models, it is essential that concepts are logically related in the bibliometric theory. Necessary and sufficient preconditions need to be stated to ensure that causes and not consequences are identified. For example, is author productivity a function of field affiliation, or is it the other way around?

It is also important to determine how author productivity might be changed by internal motivations, outside influences or manipulation. It might be assumed that tenure and promotion requirements for college and university faculty influence the degree to which individuals produce manuscripts for publication. It would be interesting to investigate the influence of such requirements on author productivity. Such a study is but one method to inject the dynamics of change into the multivariate model discussed earlier. Another test of this hypothesis would be to compare publication patterns of academic librarians who have faculty status (and might be expected to publish) with those who do not have faculty status. Even at the descriptive level, this could have an influence on the exponent associated with the Lotka distribution. External factors could also influence publication patterns of authors. Again, librarianship could be used in the investigation of this hypothesis. Many new library journals and new library publishers of monographs were formed during the past five years. It might be hypothesized that these external events have influenced the rate of author productivity in librarianship over the past decade.

The Bradford distribution (or Law of Scatter) groups journals and articles to identify the number of periodicals relevant to a particular subject. Its computation is based on the total number of articles published by the journals in a particular subject area. A constant is then computed for that subject area, which is used to determine the percentage of total coverage by various numbers of journals in a field. One formula for this is:

$$R(n) = N \log n/s \quad (1 \leq n \leq N)$$

where

$R(n)$ = total number of journal articles
N = total number of journals
s = a constant (specific to a subject area).[27]

For example, Brookes applies this formula to a scientific literature which yielded a total of 2000 articles from 400 journals. The results indicate that 40 percent of the articles are contained in 5 percent of the journals. Further, 80 percent of the articles are contained in 37 percent of the journals.[28] A core of journals is thus identified which could be used to select the essential journals for a special collection.

Originally, Bradford had studied articles and journals to improve abstracting services. He was concerned about the statistical distribution he identified, and Fairthorne reports on this: "Though in public and, rather ambiguously, in private Bradford tended to belittle this finding, he did make use of it. His private conversations gave me the impression that he was sure...that he had not enough evidence or explanation to sustain it in public debate."[29] Others have since affirmed that there is enough evidence to support Bradford's statistical distribution and to link it to a general bibliometric distribution.[30] Brookes cites numerous uses of a Bradford bibliograph: items borrowed from a library, users ranked by number of items they borrow, number of items cited (using a nonrestrictive Bradford-Zipf distribution), and the index terms assigned to documents.[31] These uses of a Bradford distribution have value for library decision-making, since the distribution allows for the prediction of regularity in a variety of events. Knowledge of sources and their items (i.e., the Bradford formula) permits prediction of core collections, core users and core index terms. However, explanation is lacking which would give theoretical import to Bradford's statistical distribution.

Why, for example, do a relatively small number of journals represent the core for any given field? Is this due to human limits in handling certain quantities of information? Are many articles published to increase an author's productivity with little concern that the article be cited (or even read)?

Bradford's distribution was made more general by grouping journals according to the number of citations they receive. Using his citation indexing data base, Garfield claimed: "I can with confidence generalize Bradford's bibliographical law concerning the concentration and dispersion of the literature of individual disciplines and specialities. Going beyond Bradford's studies, I can say that a combination of the literature of individual disciplines and specialities produces a multidisciplinary core for all of science comprising no more than 1000 journals."[32] Garfield then identifies many variables besides scientific merit which might contribute to high citation frequency. It would be through the systematic study of these variables (author's reputation, circulation, number of articles published, library holding, etc.) that reasons might emerge to explain why one journal receives numerous citations while another receives very few. A similar analysis can be applied to the core users of a library. It is not enough to predict the number of core users and their amount of use; instead, the characteristics that make an individual a core user need to be identified. Do some individuals have a reading "habit" analogous to a physical addiction? Are the backgrounds of these individuals similar, and are their other information behaviors similar? Finally, it is likely that the Bradford distribution is susceptible to change. Swanson has proposed a new model for journal articles, and he advocates that authors state the reasons for citing each reference.[33] If Swanson's prototype were implemented, it might produce drastic changes in citation patterns.

All of this points to the need for a more rigorous definition of the bibliometric problem. The analyses of bibliographic information should culminate in a causal model that accounts for variabilities in such phenomena as author productivity and journal citation patterns. The line between explanation and prediction can often be confused. For example, the movement of the sun was once explained by the god Helios riding a golden chariot across the sky. Later, it was hypothesized that the sun revolved around the earth. This theory did allow for accurate predictions; for example, the Gregorian calendar was based on the theory that the sun revolved around the earth, yet the calendar errs by only one day every 3323 years. Prediction accuracy is important but it may be an artifact of empirical regularity. A bibliometric theory—if it is to be useful—must give equal emphasis to its explanatory power and its prediction accuracy.

Bibliometric Concepts and Theory Construction

There is a wide range of bibliometric concerns beyond author productivity or journal citation patterns, and these varied interests may create problems in the development of a unified theory. This will be examined in more detail after related bibliometric topics are identified.

One area often included in bibliometric reviews is Zipf's law. It is a statistical distribution based on a hyperbolic curve which "states that, if words are ranked according to their frequency of occurrence (f), the *n*th ranking word will appear approximately k/n times where k is a constant, or $f(n) = k/n$."[34] Zipf's law has much potential for the descriptive evaluation of subject authority files and related aspects of indexing.

Other major areas of interest which could fall within bibliometrics include the half-life rates to assess the currency of a literature and impact factors to evaluate the importance of journals. Burton and Kebler studied the half-life of different scientific literatures to identify the obsolescence rate of references in journal articles.[35] For example, physics literature has a half-life of 4.6 years (i.e., one-half of all references in journal articles were dated within the last 4.6 years), while chemistry has a half-life of 8.1 years. Another view of obsolescence is to relate it to the growth of a literature: "the faster the rate of growth, the less is the scatter and the more rapid the obsolescence."[36] Closely related to half-life is Price's index to assess the hardness of journals.[37] Those journals with very recent references are considered to be at the research front as a hard science. Those journals with references to more retrospective materials are considered less hard, less scientific. For example, physics journals contain the highest percentage of references to materials published in the past five years (over 60 percent), while some English literature journals only have 10 percent of their references dated in the past five years.

Garfield developed a journal's impact factor as the number of citations a journal receives divided by the number of articles published in a given time period.[38] Narin developed influence weights as the total number of citations to a journal divided by the total number of references from a journal (excluding self-reference and self-citation).[39] Although these measures are used to evaluate journals, they can also be extended to evaluate authors by the number of citations individuals receive. Meadows gives an account of the uses of such citations to assess an author's reputation and importance.[40]

These various measures employ different units of analysis, and this creates a problem of generality across bibliometric studies. McGrath gives an excellent treatment of the unit of analysis problem as it relates to collection development.[41] He distinguishes among the objects studied (i.e., the unit of analysis), the attributes of those objects (i.e., the variables), and the appropriate levels of theoretical generality. These distinctions are applicable to the bibliometric problem. For example, if author productivity is the area under investigation, then authors are the unit of analysis and their publications are the dependent variable. The explanatory or independent variables would be those that influence an author to contribute to the publication process (as discussed earlier in relation to the Lotka distribution). This same unit of analysis—authors—would be used in investigations of author citation rates to assess the significance of an individual's contributions. The number of times an author is cited or the author's average number of citations per journal article might serve as the dependent variable. The independent variables could come from measures of collegial support, number of professional papers delivered at meetings, individual's influence on students, and the individual's personal characteristics. Author productivity and author importance could be investigated in the same study because they share the same unit of analysis. However, this is not true for the other areas of bibliometrics.

Journal citation patterns shift the unit of analysis from individuals to journals. The dependent measure might be currency of references or number of citations the journal receives from other publications. The independent variables could encompass the journal's refereeing process, manuscript acceptance rate, number of articles the journal publishes, some rating of the journal's prestige, and number of library or individual subscriptions. Of course, numerous independent variables could be posited to explain the number of citations a journal receives. But this unit of analysis—the journal—changes if the Zipf distribution is under investigation.

Zipf's law drops the unit of analysis to the word. A dependent measure might be the frequency of the word and the independent variables could include measures on the fundamental structure of language. Other explanatory variables might be the various principles associated with vocabulary control or the structure of indexing terms. These independent variables are subject to manipulation to determine the effect they may have on word frequencies. Thus, bibliometrics spans three major units of analysis: authors, journals and words. There is a fourth unit—subject or discipline—not covered here, but it is implied in the work of those who distinguish the differences across fields or disciplines (e.g., the behaviors of the literatures associated with the humanities versus the literature of the social sciences versus the sciences).[42] Much of this research has focused on the literatures of the scientific disciplines.

Since independent variables are grouped into conceptual areas the interrelationships of which become the theory, the unit of analysis is critical to the generality of the results. It is unlikely that research results would ever be generalized beyond the unit of analysis. It could prove impossible to generalize a common theory from studies of individuals and studies of journals. At best, two middle-range theories might be developed which could suggest hypotheses for a single, third area of investigation. This hope of a unified theory has plagued other professions, and it is doubtful that bibliometrics can surpass the barrier created by multiple units of analysis. Instead, it might be more productive to split the ill-defined field of bibliometrics into separate components where the unit of analysis is consistent and results can be generalized across studies.

The various bibliometric models proposed here will need to pay close attention to the issue of external validity. The models need to be more than explanatory (i.e., explaining a large proportion of the variability in the dependent measure); indeed, the models will have to prove their worth by making actual predictions using new cases. This allows for the importance (or weight) of each variable in the model to be tested in a rigorous manner. It provides proof that the theory works with new data in real situations. It also assures that hypothesized nonlinear relationships among the independent variables do, in fact, contribute to explaining the variability in the dependent measures.

Finally, bibliometrics has much to offer the library and information field. The work of the past—by Lotka, Bradford and Zipf—is valuable in helping librarians assess patterns of authorship (for cataloging rule changes), identifying core collections (for collection management), and designing better retrieval systems (for authority control). However, the continued emphasis on the similarities of the bibliometric statistical distributions is not regarded here as a fruitful endeavor. The long-term benefits of bibliometrics will begin to emerge when attention is directed toward causal explanations of bibliographic phenomena. At that point, bibliometrics will again offer practical benefits to libraries.

References

1. Pritchard, Alan. "Statistical Bibliography or Bibliometrics?" *Journal of Documentation* 25(Dec. 1969):348.

2. ________ . "Computers, Statistical Bibliography and Abstracting Services," 1968. (unpublished); and Raising, L.M. "Statistical Bibliography in Health Sciences." *Bulletin of the Medical Library Association* 50(July 1962):450, 461. Cited *in* Pritchard, "Statistical Bibliography," p. 349.

3. Simon, Herbert R. "Why Analyze Bibliographies?" *Library Trends* 22(July 1973):3-8; and Nicholas, David, and Ritchie, Maureen. *Literature and Bibliometrics*. London: Clive Bingley, 1978.

4. Hjerppe, Roland. *A Bibliography of Bibliometrics and Citation Indexing and Analysis*. Stockholm: Royal Institute of Technology Library, 1980.

5. Pritchard, Alan, and Wittig, Glen. *Bibliometrics: A Bibliography and Index* (1874-1959), vol. 1, Watford, Eng.: ALLM Books, in press.

6. Fairthorne, Robert A. "Empirical Hypbberbolic Distributions (Bradford—Zipf—Mandelbrot) for Bibliometric Description and Prediction." *Journal of Documentation* 25(Dec. 1969):319-43; Price, Derek de Solla. "A General Theory of Bibliometric and Other Cumulative Advantage Processes." *Journal of the ASIS* 27(Sept.-Oct. 1976):292-306; and Bookstein, Abraham. "The Bibliometric Distributions." *Library Quarterly* 46(Oct. 1976):416-23.

7. Narin, Francis, and Moll, Joy K. "Bibliometrics." In *Annual Review of Information Science and Technology*, edited by Martha E. Williams, p. 45. Vol. 12. Washington, D.C.: American Society for Information Science, 1977.

8. Fairthorne, "Empirical Hypberbolic Distributions," p. 322.

9. Rapoport, Anatol. "Rank-Size Relations." In *International Encyclopedia of Statistics*, edited by William Kruskal and Judith Tanur, p. 851. New York: Free Press, 1978.

10. Ibid., pp. 847-54; Price, "A General Theory"; and Bookstein, "Bibliometric Distributions."

11. Fairthorne, "Empirical Hyperbolic Distributions," p. 321.

12. Carnap, Rudolf. *Philosophical Foundations of Physics*. New York: Basic Books, 1966, p. 228.

13. Ibid., p. 230.

14. Fairthorne, "Empirical Hyperbolic Distributions," p. 332.

15. Price, "A General Theory."

16. Rapoport, "Rank-Size Relations," p. 851.

17. Ibid., p. 853.

18. Hill, Bruce M. "Zipf's Law and Prior Distributions for the Composition of a Population." *Journal of the American Statistical Association* 65(Sept. 1970):1230.

19. Line, Maurice B., and Sandison, Alexander. "Practical Interpretation of Citation and Library Use Studies." *College & Research Libraries* 36(Sept. 1975):393-96; and Line, Maurice B. "Rank Lists Based on Citations and Library Uses as Indicators of Journal Usage in Individual Libraries." *Collection Management* 2(Winter 1978):313-16.

20. Broadus, Robert N. "The Applications of Citation Analyses to Library Collection Building." *Advances in Librarianship*, vol. 7, edited by Melvin J. Voigt and Michael H. Harris, pp. 299-335, New York: Academic Press, 1977.

21. *See* Moll, Joy K. "Bibliometrics in Library Collection Management: Preface to the Special Issue on Bibliometrics." *Collection Management* 2(Fall 1978):195-98.

22. Hill, Bruce M. "The Rank-Frequency Form of Zipf's Law." *Journal of the American Statistical Association* 69(Dec. 1974):1025.

23. Narin and Moll, "Bibliometrics," p. 46.

24. Merton, Robert K. *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago: University of Chicago Press, 1973; and Zuckerman, Harriet. *Scientific Elite: Nobel Laureates in the United States*. New York: Free Press, 1977.

25. Lindsey, Duncan. *The Scientific Publication System in Social Science*. San Francisco: Jossey-Bass, 1978, p. 89.

26. Mischel, Walter. "Toward a Cognitive Social Learning Reconceptualization of Personality," *Psychological Review* 80(July 1973):252-83.

27. Brookes, B.C. "Numerical Methods of Bibliographic Analysis."*Library Trends* 22(July 1973):26.

28. Ibid., p. 27.

29. Fairthorne, "Empirical Hyperbolic Distributions," p. 333.

30. Price, "A General Theory"; and Bookstein, "Bibliometric Distributions."

31. Brookes, "Numerical Methods of Bibliographic Analysis."

32. Garfield, Eugene. "Citation Analysis as a Tool in Journal Evaluation." *Science* 178(3 Nov. 1972):476.

33. Swanson, Don R. "Information Retrieval as a Trial-and-Error Process," *Library Quarterly* 47(April 1977):128-48.

34. Narin and Moll, "Bibliometrics," p. 46.

35. Burton, Robert E., and Kebler, R.W. "The 'Half-Life' of Some Scientific and Technical Literatures." *American Documentation* 11(Jan. 1960):18-22.

36. Brookes, "Numerical Methods of Bibliographic Analysis," p. 34.

37. Price, Derek de Solla. "Citation Measures of Hard Science, Soft Science, Technology, and Nonscience." In *Communication Among Scientists and Engineers*, edited by Carnot E. Nelson and Donald K. Pollock, pp. 3-22. Lexington, Mass.: Heath Lexington Books, 1970.

38. Garfield, "Citation Analysis," pp. 471-79.

39. Narin, Francis. *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity*. Cherry Hill, N.J.: Computer Horizons, 1976. (PB 252 339)

40. Meadows, Arthur J. *Communication in Science*. London: Butterworths, 1974.

41. McGrath, William E. "Circulation Studies and Collection Development." *Collection Development in Libraries*, edited by Robert D. Stueart and George B. Miller, Jr., pp. 373-403. Greenwich, Conn.: JAI Press, 1980.

42. Lindsey, *Scientific Publication System*; Price, "Citation Measures of Hard Science"; and Garvey, William D. *Communication: The Essence of Science*. Oxford: Pergamon Press, 1979.