Bradford's Law: Theory, Empiricism and the Gaps Between

M. CARL DROTT

NATURAL LAWS DESCRIBE PATTERNS which are regular and recurring. The scientific point of a law is twofold. First, a concrete statement of a law may give give us the ability to better predict events or to shape our reactions to them. Second, a physical law may help in the development of theories which explain why a particular pattern occurs. Natural laws therefore are of interest because they offer the opportunity for empirical application and for theoretical understanding. On the other hand, the ability to articulate a law does not automatically guarantee either empirical or theoretical advances.

Bradford's law begins with a regularity which is observed in the retrieval or use of published information. Broadly speaking, this regularity is characterized by both concentration and dispersion of specific items of information over different sources of information. Thus, for a search on some specific topic, a large number of the relevant articles will be concentrated in a small number of journal titles. The remaining articles will be dispersed over a large number of titles. Throughout the remaining discussion, journal articles will be used to represent the items retrieved and journals will be the sources. This is in keeping with most of the Bradford's law literature, although there is clear evidence that similar patterns occur for other kinds of items and sources.

The literature on Bradford's law incorporates both theoretical and empirical aspects. These aspects are each coherent and developing areas of scientific inquiry. Confusion arises, however, when the two aspects

M. Carl Drott is Associate Professor, School of Library and Information Science, Drexel University, Philadelphia.

become mixed. This mixing occurs in the normal course of scholarship. Authors with empirical data quite properly speculate on what might be implied in terms of theory. Writers developing theoretical models offer empirical interpretations as a way of making the abstract more concrete. It is important for readers and future researchers to separate clearly the knowledge developed in each aspect from the many unanswered questions which separate theory from empiricism.

Theoretical Development

The fundamental question in the theoretical study of Bradford's law is this: What is the nature of the underlying probabilistic events which aggregate to create the regular pattern of dispersion of articles over titles? As a first step toward solving this difficult (and as yet unsolved) problem, it is necessary to have a mathematical description of the pattern whose appearance we are trying to explain. The first statement of this mathematical formula came from S.C. Bradford.[1] He examined all of the journal titles contributing to a bibliography on applied geophysics. Bradford discovered that he could divide the titles into three groups, such that each group of titles contributed about the same number of articles. Starting with the titles which contributed the most articles, he divided the articles into three roughly equal groups:

The first 9 titles contributed 429 articles.
The next 59 titles contributed 499 articles.
The last 258 titles contributed 404 articles.

The value of this arrangement lies in the number of titles it takes for each one-third of the articles. In this case, Bradford discovered a regularity in calculating the number of titles in each of the three groups:

9 titles
9 × 5 titles (equals 45 titles)
9 × 5 × 5 titles (equals 225 titles)

Just as the three groups of articles were not quite equal in size, this formulation does not quite give the observed number of titles. This arrangement does have a very special regularity. There is a "core" of nine titles which contributes one-third of all the articles. In order to get the second third of the articles (that is, to add the same number of articles already found), one needs to search five times as many titles (5 × 9). To find the last third of the articles (again, to add the same number of articles as found in the "core" titles), one must search five times again (9

× 5 × 5) as many titles. Thus, to show title groups contributing an equal number of articles, one could write:

$$9 : 9 \times 5 : 9 \times 5^2$$

Recognizing that the size of the core (9) and the multiplier (5) might be different for other searches, we divide the groups by nine and replace the multiplier with a variable. This gives groups of titles with sizes:

$$1 : a : a^2$$

where each of the three groups of titles contributes the same number of articles.

This is the first theoretical statement of Bradford's law. Note that while it was founded on empirical observation, it is not derived strictly from the data. (As noted above, the data do not quite fit the law either in the exact number of articles in each group or in matching the calculated number of titles to the observed number.) As a statement of a natural law this formulation has several shortcomings. The most serious problem is that the phenomenon is described in terms of groups of journals. These rather large aggregations of titles seem to be an artifact of the statement of the law. That is, it appears that the dispersion of articles over ranked titles is mathematically regular rank by rank rather than being regular only for groups. There is also no hint in the formula or its derivation as to what kind of underlying probabilistic process creates this scattering. Bradford's formulation also leaves unanswered questions for those working with empirical data. How does one establish the size of the core? What is the "best" value of *a* for any particular set of data (recognizing that, as above, no value of *a* fits the observations exactly)? These questions are indicative of the gap that arises between empirical and theoretical consideration of the phenomenon.

Work on clarifying and refining the theoretical statement of Bradford's law was undertaken by B.C. Vickery,[2] M.G. Kendall,[3] F.F. Leimkuhler,[4] and others. The most profound impact on the theoretical foundation of Bradford's law has come from the efforts of B.C. Brookes.[5]

Brookes began with Bradford's ratios as portrayed above. Drawing on the work of Vickery, he derived a formula which did not depend on groupings of journal titles. The formula was this:

$$R(n) = k \log (n)$$

where:

n is the rank of each journal

In other words, the journal contributing the most articles has a rank of 1, the second most productive title has a rank of 2, and so on. In assigning ranks, every title is given a rank. In the case of ties (titles contributing the same number of articles), ranks are arbitrarily assigned to the tied journals.

R(n) is the total number of articles contributed by the first *n* journals. The value of R(1) is simply the number of articles contributed by the top title. The value of R(2) is the sum of the number of articles contributed by the first journal plus the articles contributed by the second-ranked title.

k is a constant which may be different for each search. It is related to the document collection.

Note that this formula can be used to calculate the number of articles contributed by a journal at any rank. For example, the number of articles contributed by the fifth-ranked journal is simply R(5) - R(4) (the total number of articles contributed by the first five titles minus the number of articles contributed by the first four titles).

This formulation of Bradford's law allows us to use much greater mathematical power in the search for an understanding of the theoretical aspects of the problem. One way of seeking this understanding is to consider what the equation implies about the real world. If predictions made from theory are obviously false, then we know that there is some error. Either the theory must be changed, or there must be some restrictions included as to exactly what phenomenon is being described. Note that the converse is not true. The fact that the theory does fit the world does not actually prove the truth of the theory.

Brookes used the following approach in refining his formulation. He considered the predictions which the formula made when the search retrieved a very large number of articles. In such a situation, the formula required the number of articles contributed by each of the top-ranked journals to grow very large. However, we know that there must be a limit to the number of articles on a topic which any single journal can publish even if it deals with nothing but the topic. Further, there are a number of empirical studies which show that the number of articles contributed by the top-ranked journals is not as high as the formula would predict. Strictly speaking, the prediction from the formula is too low for the first journal and too high for the remaining most-used journals. In fact, for some data sets the formula predicts that the number of articles contributed by the top-ranked titles will be negative.

In order to account for this disparity, Brookes modified the formula to include another constant, *s*.

$$R(n) = k \log (n/s)$$

He also imposed the limitation that this statement of Bradford's law may not hold for the most frequently appearing titles in a data set. This modification can be viewed as a speculation on the fundamental theoretical question. That question asks the underlying reason for the observed regularity. This modification, in essence, says that the underlying process which creates the regularity may be different from the process which causes the top-ranked titles to diverge from regularity. In other words, the behavior of the top-ranked journals may present a different theoretical problem than the pattern of the remaining titles.

There is another problem in accommodating the mathematical form of Bradford's law to the observed data. In this case, the issue involved those titles which contribute only a few articles (or a single article) each. Empirical data show that there are not as many of these little-used sources as the theory would predict. If the formula is correct, then the total number of titles found must be exactly the value of k. In practice, observed searches fall short of this number.

The data on little-used titles again raise a problem for theorists: either to modify the statement of the law or to reject the empirical data. Rejecting the data in this case means assuming that the observed searches are incomplete. Realistically, however, many of the searches are well and painstakingly done. It is hard to imagine how they could be made more complete.

Theorists have chosen to accept the mathematical formula and reject the empirical data. The reasons for this choice illustrate an important aspect of the difference between theory and empiricism. The important factor to theorists is that the mathematical form of Bradford's law as stated above is very "agreeable" in a mathematical sense. In its present form, Bradford's law can be related to other mathematical models of dispersion. These models include the gamma, Poisson, and binomial distributions. These other distributions have been extensively studied. The scattering phenomena which these distributions have been shown to describe seem related to bibliometric scattering. Thus, in rejecting the empirical data, theorists are not saying that they believe that searches are incomplete or that k truly predicts the true number of titles that will be found. Theorists are instead saying that they believe that the advancement of understanding lies in the study of certain mathematical forms. The question of conformity to empirical data is seen as less important in this situation.

The decision not to alter the mathematical form of Bradford's law has another advantage in the development of theory. The advantage lies in the fact that the formula is still assumed to apply to the titles contributing only a few articles. To a librarian, the journals which contribute only an occasional article on a topic of interest are of much less importance than those which regularly have many relevant articles. Theoretical development requires a slightly different perspective.

Consider the way in which the literature on a new topic develops. Initially, no journals have any articles on the subject. Then as the field develops, some journals publish their first article. Of all the journals that publish a first article on the subject, some fraction will publish a second article. Similarly, those journals publishing any number of articles are a fraction of those titles which published one fewer than that number of articles. Viewed in this manner, the publication of a small number of articles is a step toward publishing a greater number. This line of reasoning makes it desirable not to exclude journals contributing only one or two articles from the development of the theory. In a sense such items are the base on which the distribution is built.

Brookes noted that in this progression, only those journals which have succeeded in publishing at some level can have a chance of rising above that level. Thus, since the competition diminishes, each remaining journal stands an even better chance of attracting articles. This kind of "success breeds success" pattern was articulated by Derek de Solla Price[6] in his cumulative advantage model. This model has the possibility of adding to our theoretical understanding of Bradford's law. It also offers a broader understanding of other related bibliometric distributions. Thus, in scope, this theoretical development goes beyond Bradford's law to a much broader class of probabilistic phenomena.

Empirical Development

The fundamental question in the empirical study of Bradford's law is this: What are the implications of the observed pattern for the provision of user service? This involves two aspects: prediction and evaluation. Prediction could tell what titles would be useful or how users would behave. Evaluation could provide a theoretical standard against which retrieval or acquisition could be measured.

Empirical studies generally begin with a rank-frequency table. The steps in the creation and interpretation of such a table have appeared elsewhere.[7] Typically, such a table lists each rank, the number of articles contributed by the journal of that rank, a cumulative frequency corresponding to the variable R(n), and a cumulative percentage. From an empirical point of view, the cumulative percentage of articles is the most important. The pattern is that a high percentage of the articles comes from a very small number of journals. At this point any knowledgeable librarian can nod in agreement. Good practice dictates that the most-used titles must be identified and their availability assured. On the other hand, there are a large number of titles with low usage. Only the largest budget could justify holding them all. Yet, it is clear that access must be provided.

The discussion above is better classed as conventional wisdom than as exploitation of a natural law. The challenge (as yet unmet) of empirical studies is to find a way of using quantitative regularity to make decisions which are more precise than simple intuition would provide.

Before we can say much about using Bradford's law, we must have some way of knowing if a set of data conforms to the law. This immediately raises problems. In every kind of goodness-of-fit test we need to have some source of predicted values against which to judge our data. Thus, we must ask the question: What is Bradford's law? The usual answer is that it is the formula for R(n) given earlier. But this is not completely rational. As discussed above, the formula is known to be in disagreement with empirical observation. Further, the formula excludes the most-used titles, which in many actual situations may be the most important. This exclusion is complicated by the fact that exactly how many titles are to be excluded is undefined. This number is usually determined by the process of inspection, a rather arbitrary procedure.

In spite of the problems, the formula given above is generally taken as the source of expected values. This means that one must obtain values for *k* and *s*, the two constants in the equation. These are obtained by recognizing that if ideal data were plotted with one axis for R(n) (cumulative articles) and the other for log (n) (log rank), the result would be a straight line. The variable *k* and *s* represent the slope and intercept, respectively, of that line. The usual process for obtaining these values follows. First, the data are plotted on semilogarithmic graph paper. Next, a straight line is drawn through some central portion of the curve. This offers the investigator an arbitrary choice as to how much of the data to use and exactly what straight line "best" fits those data. The value of the slope (*k*) is determined for the line. This is often done by using only two points, thus introducing further arbitrariness. The intercept (*s*) is obtained either by graphical extrapolation or by using the slope and a point on the line.

There is an alternate procedure to determine the constants. This method uses linear regression on the data (or an arbitrarily selected part of the data). This approach has the advantages of being more replicable and of using more of the data. The disadvantage is that rank, a clearly ordinal measure, is treated as if it were on an interval scale. Such an assumption is not unique to this application, but it must give the thoughtful researcher reason to pause.

With the constants determined, expected values of R(n) can be calculated for each rank. Next, a statistical test must be used to compare the observed and expected values. This raises another difficulty. On the one hand, we know that because of the assumptions made, we do not expect an exact fit. On the other hand, the ranking process imposes an order on the data so that there will always be some degree of association between R(n) and *n*.

The most frequently used test in this situation is the chi-square test. This requires an arbitrary grouping in order to avoid cells with small numbers. A greater problem is the tendency of chi-square to find significant differences whenever the sample size is large.[8] This is a special problem in this situation, since we know that some difference between expected and observed must exist.

An alternative measure is Pearson's correlation. This measure of variance reduction does not provide an answer as to whether a hypothesis should be accepted or rejected. Thus, the rigid arbitrariness of the chi-square test is replaced with the arbitrary opinion of the investigator. Correlation also suffers from the drawbacks of regression analysis on which it is based. (Note that because the data are ranked, the test for the significance of a correlation is meaningless.)

Some other measures to test for conformity to Bradford's law have been proposed. The Kolmogorov-Smirnov test has been proposed as an alternative to chi-square.[9] More experience with this test will be needed before its worth can be evaluated. Another, more informal approach is to calculate values of the intercept (*s*) for a number of observed data points. Close agreement of these values is taken to indicate a Bradford-type distribution.

The statistical problems of identifying a Bradford distribution are compounded when comparing several sets of empirical data. In this case, the question is not only the form of the distribution, but also whether the distributions are the same. One problem is that the constants will produce a shift in the cumulative percentages for each rank. The nature of this shift is complex because both the number of articles and the number of titles are shifting. There seems to be no accepted statistical test for this situation.

Even if the sample sizes are the same, it is still difficult to determine if two data sets should be considered identical within the limits of sampling error. This problem frequently arises when samples are taken in the same situation but at different times. Some of the variation in the rankings of titles will be due to sampling error. But changes in rank may also reflect real changes in the use of a title. The sample sizes required to resolve this issue are very large indeed. For example, Brookes has calculated that to achieve a 95 percent confidence level that two adjacent titles should not reverse their order, a sample size of several thousand—if the titles are high (e.g., 5 or 6) in the ranking—is required.[10] The resolution of lower-ranked pairs requires much larger samples (tens or hundreds of thousands). Consideration of these sample sizes should make any researcher cautious in accepting the accuracy of empirical data.

The Gap Between

The title of this article alludes to a gap between theoretical studies of Bradford's law and empirical research. The gap is this: none of the variables which characterize the empirical situation have been shown to relate to the theoretical model. These include variables which describe the field or topic being researched, the way the search is conducted, the specific needs of the user, or the characteristics of the collections involved. This is a rather peculiar situation. Anyone with practical experience in information retrieval recognizes that these parameters are important in providing high-quality service. It is almost contra-intuitive to find that none of these variables are reflected in the theoretical study of Bradford's law.

There is an important limitation to the gap described above. It is well known that the size of the set of retrieved items (in terms of both total articles and total journal titles) is related to the theoretical model. The number of articles is strongly related to the slope (constant k in the equation), and the number of titles is somewhat related to the intercept (constant s). Thus, any aspect of the empirical situation which affects these values will have a tie to the theoretical model. For example, the generality or specificity of the topic (for a given field) may affect the number of items retrieved. In such a case, the topic breadth will seem to affect the model. In fact, this effect is related to a change in the number of articles and titles, not to intellectual characteristics of the topic.

This relationship leads to some very odd conclusions for the unwary investigator. For example, Pratt has proposed a measure of the degree to which articles in a particular field are concentrated within the literature.[11] The claim is made that this index can be used with

Bradford-type data. (The claim is actually made for Zipf-type data, a mathematically identical distribution.) But Pratt's index depends on the number of titles in the sample. Consider two sets of data on exactly the same topic: for example, Lawani's searches on tropical agriculture for one year and four years.[12] The Pratt index, affected by sample size, would lead to the conclusion that tropical agriculture is a more concentrated field than tropical agriculture.

A failure to recognize that data are subject to sampling error can also produce meaningless "applications" of Bradford's law. For example, Goffman and Morris propose that circulation samples from a journal collection be used to predict the distribution of use for the next year.[13] They propose a one- to three-month sampling period and give an example with a sample size of 876. They claim a "core" of eleven titles. They do not actually make a prediction or test it. According to Brookes, the appropriate sample size for this situation is about 25,000. Given the huge undersampling proposed, the Goffman and Morris study is better classed as an application of common sense rather than any use of Bradford's law.

Aside from the misuse of Bradford's law, the question arises as to whether the gap between theory and practice is simply due to the fact that more research findings are needed. This corresponds to the hypothesis that empirical variables (those which characterize the intellectual dimensions of retrieval) can be incorporated into the theoretical model. The alternate hypothesis is that the role of the empirical variables is only to define those situations for which the model can be expected to hold. In this case, the empirical variables are constraints or limits but not an actual part of the theoretical model. One area of empirical data which may shed light on this gap is the behavior of the most popular journal titles. In the discussion of theoretical development earlier in this paper, it was noted that in some empirical situations the most frequently occurring titles contribute fewer articles than would be expected. A proposed interpretation of this divergence is that the top journals become "saturated" with articles on the topic. This explanation seems very reasonable, but has never been substantiated.

If empirical variables such as the size, areas of specialization, and editorial policies of the top journals have an effect, then it should be possible to relate different levels of saturation to different empirical circumstances. This would serve, finally, to tie the theoretical model to empirical parameters.

Summary

The literature on Bradford's law presents the casual reader with a number of pitfalls. The first problem is to distinguish theoretical from empirical research. Theoretical work is aimed at understanding a random probabilistic process. To this end, assumptions are made which aid mathematical manipulation. Empirical studies concentrate on describing the world from a practitioner's point of view. In these studies the descriptive qualities of the data are more important than the statistical aspects. A second problem is the large number of "marginal" claims in the literature, that is, claims which are clearly speculative or are simply unsupported. Some of this writing is not intended for acceptance without further study. Other articles are simply weak scholarship. In both cases the reader must decide what to reject.

Between theory and empiricism lies a gap. This gap is the fact that at present, the intellectual richness of real situations is not represented in the mathematical austerity of the theoretical equations. It remains to be seen if this gap can be bridged by further research.

Overall, Bradford's law represents an elusive phenomenon. On one hand, it is easy to observe in real situations and can be represented with a fairly simple mathematical formula. On the other hand, Bradford-type data resist statistical testing, and the model fails to reveal the underlying process which "causes" the distribution. In any case, the wise reader will examine any study of Bradford's law closely before rushing to believe more than is actually stated and supported.

References

1. Bradford, Samuel C. "Sources of Information on Specific Subjects." *Engineering* 137(26 Jan. 1934):85-86; and ———. *Documentation.* Washington, D.C.: Public Affairs Press, 1950.

2. Vickery, B.C. "Bradford's Law of Scattering." *Journal of Documentation* 4(Dec. 1948):198-203.

3. Kendall, M.G. "The Bibliography of Operational Research." *Operational Research Quarterly* 11(March/June 1960):31-36.

4. Leimkuhler, Ferdinand F. "The Bradford Distribution." *Journal of Documentation* 23(Sept. 1967):197-207.

5. Brookes, Bertram C. "The Derivation and Application of the Bradford-Zipf Distribution." *Journal of Documentation* 24(Dec. 1968):247-65; ———. "Bradford's Law and the Bibliography of Science." *Nature* 224(6 Dec. 1969):953-56; and ———. "Obsolescence of Special Library Periodicals: Sampling Errors and Utility Contours." *Journal of the ASIS* 21(Sept.-Oct. 1970):320-29.

6. Price, Derek de Solla. "A General Theory of Bibliometric and Other Cumulative Advantage Processes." *Journal of the ASIS* 27(Sept.-Oct. 1976):292-306.

7. Drott, M. Carl, et al. "Bradford's Law and Libraries: Present Applications—Potential Promised." *ASLIB Proceedings* 31(June 1979):296-304.

8. Mosteller, Frederick, and Wallace, David L. *Inference and Disputed Authorship: The Federalist.* Reading, Mass: Addison-Wesley, 1964.

9. Brookes, Bertram C. "Theory of the Bradford Law." *Journal of Documentation* 33(Sept. 1977):180-209.

10. Ibid.

11. Pratt, Allan D. "A Measure of Class Concentration in Bibliometrics." *Journal of the ASIS* 28(Sept. 1977):285-92.

12. Lawani, S.M. "Periodical Literature of Tropical and Subtropical Agriculture." *Unesco Bulletin for Libraries* 26(March-April 1972):88-93; and ________. "Bradford's Law and the Literature of Agriculture." *International Library Review* 5(July 1973): 341-50.

13. Goffman, William, and Morris, Thomas G. "Bradford's Law Applied to the Maintenance of Library Collections." In *Introduction to Information Science,* edited by Tefko Saracevic, pp. 200-03. New York: Bowker, 1970.