December 2003

Bradford's Distribution, the 80/20 Rule, and Patterns of Full-Text Database Use

Steve Black

The College of Saint Rose, blacks@mail.strose.edu

Amy Sisson

The College of Saint Rose, sissona@strose.edu

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>

Part of the [Library and Information Science Commons](#)

Recommended Citation

Black, Steve and Sisson, Amy (2003) "Bradford's Distribution, the 80/20 Rule, and Patterns of Full-Text Database Use," *Against the Grain*: Vol. 15: Iss. 6, Article 7.

DOI: <https://doi.org/10.7771/2380-176X.4198>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

Making Haste Slowly: from page 18

their nature and not the least of that is a result of the political fact that many parties are coming to the table with different objectives. Fine, so we will start with a set of least common denominators. Let's work to make **COUNTER** the best least common denominator we can and build on that foundation based on our congruent interests.

Sometimes these incentives will lead us to cooperation and sometimes to contention but, in any standards body, this kind of contention occurs, particularly as we have seen in Internet standards bodies where commercial and not-for-profit interests have sometimes clashed sharply. The goal is better data and if we work toward that end in a disciplined fashion, we will have a centralizing force to overcome the centrifugal forces these disparate objectives will generate.

OTOH, Let's Keep the Powder Dry

There are generally two kinds of uses for data: decision support or research into relationships leading to understanding and prediction. In a practical field like the library field, there should be overlap but the record in the field on analyzing data of any kind is dreadful. Library science is not in any real sense a science.

That said, it is impressive what is going on in the measurement of use of electronic materials and how people interact with them. In fact, there is probably a sufficient body of material for a useful bibliographic essay of the scope of those found in the *Annual Review of Information Science and Technology*. I think we actually know a fair amount if we had everything collected in one spot... here is a job for evidence-based librarianship. I have a sense we know a lot more than we think we know if it were all collected and assessed.

While there are articles based on data gathered in the libraries from direct measurement, most articles one sees use vendor-supplied data. The kind of data gathering outside vendor-derived data is taking place in institutions where there is technical expertise. OhioLINK <http://www.ohiolink.edu/>,³ and the **University of Pennsylvania**⁴ are two I know about but I am sure there are others. In these institutions, the actual streams of data are being analyzed and these institutions are not depending on vendors for all their data. I think these kinds of independent initiatives are a vital development.

Here is a place where individual use of materials may finally give us a means to measure the effect of libraries on people directly while respecting the privacy of the subjects. Here, too, data that will not be available from the **COUNTER** process for whatever reason may be investigated. Maybe you won't be able to get this or that important number from the vendors but maybe you can get ideas of what those values would be from other sources or perhaps these institutions will make software available to do the work locally.

What next?

There is an interesting mix of factors in providing digital information that will affect the future in the near term but this is the beginning

of a revolution that will take centuries to work out. That fact means that there is bound to be more change and more difficulties with data in the future. Data for the librarian making selection decisions will start to get better quickly, though.

I blanch to say this, but on occasion the library world has relied on the kindness of strangers to collect its data and the experience has rarely been a happy one; an independent effort should be maintained. We do not want a monoculture of data on the use of electronic resources so that kind of effort is important to keep going. The independent effort should continue for another reason: **COUNTER** may fail and independent efforts outside vendor-supplied data may be what we are left with.

I think we in the library community should do everything we can to help make **COUNTER** a success and being aware of the many varying objectives of the parties involved is a way to help do that. Another way to keep the centrifugal forces in check is a countervailing force: commitment to the objective of collecting data we all trust so that we can make this new enterprise we are all engaged in work. We do not want to kill this baby in the crib.

Column Editor's Note: Thanks to **Phil Davis**, **David Goodman**, and **Chuck Hamaker** for their observations and comments. Errors are mine, of course. These opinions are my own and not those of the **National Commission on Libraries and Information Science**. — RM

References

1. **COUNTER's** Web page is at: <http://www.projectcounter.org/>. Examples of usage reports are at: http://www.projectcounter.org/code_practice.html#section4.
2. Information on **COUNTER** membership is at: <http://www.projectcounter.org/member.html>.
3. OhioLINK's Web page is at: <http://www.ohiolink.edu/>.
4. Zucca, Joe, "Find a Way or Make One: Strategies for Measuring and Analyzing the Use of Digital Information," PowerPoint slides from a talk at the ALA Annual Meeting in Toronto, June 20, 2003, <http://www.arl.org/stats/newmeas/emetrics/arljune2003.ppt> (14MB).

Bradford's Distribution, the 80/20 Rule, and Patterns of Full-Text Database Use

by **Steve Black** (Reference, Instruction, and Serials Librarian, Neil Hellman Library, The College of Saint Rose; Phone: 518-458-5494) <blacks@mail.strose.edu> and **Amy Sisson** (Neil Hellman Library, The College of Saint Rose) <sissona@strose.edu>

A well-established principle of bibliometrics is that a relatively small number of journals get the majority of use. A simple mathematical formulation of this principle is the 80/20 rule, which states that approximately 80% of journal use is concentrated in 20% of the journal titles (Trueswell 1969). The 80/20 rule may be applied to online journal use as well as print use; a recent study of consortial use of online journals in the **Academic Press IDEAL** aggregate package found that 24.3% of titles satisfied 80% of downloads (Davis 2002).

A more elaborate mathematical model for the concentration of journal use in relatively few titles is **Bradford's** distribution, stated as $1:n:n^2:n^3:n^4\dots$, which **Bradford** originally derived by examining the use of print journals within the specific disciplines of applied geophysics and lubrication (Bradford 1953, p.154). This distribution of use, also known as **Bradford's** law of scatter, is recognized as fundamental

to bibliometrics and information science, and has received significant attention since its publication (Lockett 1989). **Nisonger** (1998) provides an overview of **Bradford** distribution as it fits into the bibliometrics of journal use, and **Oluic-Vukovic** (1997) has described the development of mathematical models based on **Bradford's** law of scatter and related distributions.

To determine **Bradford's** distribution for a set of journals, the journals are ranked in use order, with the journal receiving the highest number of uses ranked first. The **Bradford** multiplier, n , is equal to how many of the next highest-ranked journals it takes to achieve the same number of uses as the highest-ranked journal. Thus, if the first journal receives 1,000 uses, and the next three ranked journals together receive approximately 1,000 uses, the **Bradford** multiplier n is 3. **Bradford's** formula then predicts that it will take the next n^2 journals, or 9 journals in this case, to receive the next 1,000 uses; the next n^3 , or 27 journals, to receive the next 1,000 uses, and the next n^4 , or 81 journals, to receive the next 1,000 uses.

During **Bradford's** day, it was pro-
continued on page 22

Bradford's Distribution, the 80/20 Rule, and Patterns ... from page 20

hibitively time-consuming to collect data on the use of large, interdisciplinary journal collections. Today, however, the use statistics that are easily generated from aggregated database packages make it possible to examine whether the **Bradford** distribution and the 80/20 rule hold for the use by college communities of interdisciplinary full-text databases.

METHOD

This study is based on use data downloaded from the administration model of **EBSCOhost**. This database package was chosen because it is the most-used package of full-text databases at **The College of Saint Rose**, it is available at some of our peer institutions, and more than two years (January 2001–present) of use data were available. In order to examine **The College of Saint Rose's** use alongside that of comparable institutions, we looked at the Websites of our peer institutions as listed in our 1996 **Middle States** accreditation report, and determined that eight of the 22 colleges offered **EBSCOhost** databases to their students. In June 2003, we sent an email requesting a use report to the appropriate librarians at those eight institutions. The email requested that the librarians log into the administration module of **EBSCOhost** and run two separate reports, one showing the number of abstracts uses and one showing the number of full-text uses, for the period from January 2001 to May 2003.

This request yielded three sets of reports in addition to **The College of Saint Rose** data, for a total of four data sets covering the period from January 2001 to May 2003. The reports were anonymous, with no institution-identifying information attached. These data sets were copied from HTML into Excel; we then excluded the abstracts and articles viewed for "unknown" titles (no ISSN) and for ERIC documents, since views of the latter are not of single titles. According to **EBSCO**, an abstract or full-text article is logged as "viewed" if it is opened, saved, or printed. A detailed description of how **EBSCO** logs transactions may be found at http://support.epnet.com/custsupport/UserDocumentation/Statistical_Documentation.doc.

RESULTS

It is not valid to compare use on a title-by-title basis from one college to the next, even if user communities are similar. Institutions subscribe to different databases via **EBSCOhost**, profile settings in the **EBSCO** administration module may affect use counts, and full-text titles may be accessed from other sources. The goal here is simply to describe the overall patterns of online journal use in **EBSCOhost** databases at the four colleges in this particular sample.

For all four colleges, the percentage of journals that achieve 80% of use is lower than the 20% predicted by the 80/20 Rule, although College D comes close, as shown in Table 1. Interestingly, the greater the overall use of **EBSCOhost** databases, the more concentrated is the portion of titles receiving 80% of the hits.

Table 1: Portion of Titles that Received 80% of Hits

Institution	Document Type	Total hits 1/2001-5/2003	% of titles receiving 80% of use
College A	Abstracts	501,768	11.87%
	Full text	326,819	12.44%
College B	Abstracts	281,351	13.70%
	Full text	109,480	13.41%
College C	Abstracts	89,071	16.91%
	Full text	80,019	17.62%
College D	Abstracts	78,127	19.12%
	Full text	83,008	18.37%

For this sample, the percentage of titles receiving zero hits was inversely related to the number of total hits. At College A, 3.82% of the titles received zero hits; the percentage receiving zero hits at Colleges B, C, and D were 3.50%, 6.20%, and 6.5%, respectively. Greater overall

database use apparently causes more titles to be viewed at least once, even though use is concentrated on a smaller percentage of titles.

It's also notable that the patterns of abstract versus full-text use are similar within institutions, but vary between institutions. In response to an email posted by one of the authors to **SERIALST**, **David Goodman** of **Princeton** wrote, "at **Princeton** we typically see a narrower distribution of use from full articles than abstracts (which I interpret as meaning that people look at abstracts from a wide range of sources, but only find the [full-text articles] from the major journals worth actually reading through)." However, in this sample, Colleges A and C show more concentration of use in abstracts than in full text, and full text use is only slightly more concentrated than abstract use at Colleges B and D.

From the ranked lists, we then applied the **Bradford** law of scattering formula, as described above, to the data. However, since use of journals did not drop off as quickly as **Bradford's** distribution would predict, all but one of the data sets would have had a **Bradford** multiplier between one and two, and therefore would not be a whole number. Therefore, we chose to cumulate uses of the top three journals in each ranked list as the one highest-ranked journal in **Bradford's** formula.

As shown in Tables 2 and 3, **Bradford's** distribution does not accurately describe the patterns of use of full-text interdisciplinary databases for either abstract or full-text use. If **Bradford's** distribution held for the use of full-text journals in **EBSCOhost** databases by the users at these four colleges, the hits for n , n^2 , n^3 and n^4 would be approximately equal. However, the number of hits for n^2 and n^3 were greater than expected. Also, titles in the middle of the ranked list received more hits than **Bradford's** distribution would predict. Overall, **Bradford's** predicted exponential increase in the number of titles required to meet the same level of use does not match this data.

Table 2: Bradford's Distribution of Hits on Abstracts

Institution	Hits for 1 (first 3 titles)	n	Hits for n	Hits for n^2	Hits for n^3	Hits for n^4
College A	28,402	7	27,507	98,403	168,451	155,183
College B	14,722	6	14,334	39,607	79,071	92,343
College C	8,263	8	8,089	19,838	28,070	24,246
College D	6,702	8	6,471	16,309	27,872	20,175

Table 3: Bradford's Distribution of Hits on Full-Text Articles

Institution	Hits for 1 (first 3 titles)	n	Hits for n	Hits for n^2	Hits for n^3	Hits for n^4
College A	13,698	4	13,333	39,550	80,927	92,038
College B	6,199	4	5,633	16,611	28,354	28,227
College C	6,577	5	6,812	15,279	22,246	22,026
College D	5,549	4	5,531	17,228	19,720	22,761

DISCUSSION

Despite the failure of **Bradford's** distribution to fit this sample, the portion of titles that received 80% of the hits supports the general principle of concentrated use in a relatively small proportion of titles. The concentration of use displayed in Table 1 appears to be due to a slower-than-predicted drop in the use of titles as one moves down a ranked list. Unlike the pattern of use of print journals in specific disciplines found by **Bradford**, the heaviest use in these interdisciplinary database packages is spread among dozens, rather than a handful, of titles.

These results may not seem surprising in that an interdisciplinary database package may contain those journals that are most fundamental to several different individual disciplines, thus creating areas of usage overlap for each discipline's most frequently used journals. The data for an institution with large numbers of both education and science majors, for instance, might show that the core journals for

continued on page 24

Why Usage Statistics Cannot Tell Us Everything, and Why We Shouldn't Dare To Ask

by Philip Davis (Life Sciences Bibliographer, Mann Library, Cornell University; Phone: 607-255-7192) <pmd8@cornell.edu>

What would you be willing to pay for a full-text download?

\$30, \$20, \$5, 25 cents?

Would you answer this question differently if you knew:

1. That it was a faculty member needing the article while preparing a million-dollar research grant application?
2. That it was a graduate student doing a comprehensive literature search for her dissertation?
3. That it was an undergraduate downloading the first three documents for an assignment due tomorrow?

Would you also answer the question differently if you knew:

1. That you were paying for an article download that you already had in your stacks?
2. That this article was downloaded first as an HTML document and then as a PDF (i.e. two downloads)?
3. That the same article was available in a competing product for a tenth of the cost?

And lastly, would you change how much you would be willing to pay if you knew:

1. That the article was actually read?
2. That only the title and abstract were browsed?
3. That someone systematically downloaded this article (along with every other article in this journal) because he wanted to create a personal archive?

While usage statistics can tell us so much about *how much* a journal or resource is being used, it cannot tell us *why* it was used, or *by whom*. Asking the *why* and *by whom* questions may be very tempting for both librarians and publishers. I will argue in this short piece that the answer to these questions have unintended consequences for library budgets.

Price Discrimination

Price discrimination is charging different prices to different customers for the same good or service. It is practiced because it results in greater profits than if a company charged a uniform price to all customers. Price discrimination is practiced at the movie theatre (age discrimination), for a haircut (gender discrimination), and for airline tickets (by charging business travelers more than vacationers). Consumers often despise price discrimination. Those who pay more don't believe that it is "fair" to pay more for the same service, and those who pay less may feel that they could have received a better deal.

The electronic publishing market is ideal for price discrimination. Infrastructure costs are very high and marginal costs are exceedingly low. The cost to create the first online copy is the publishers' greatest expense, while the cost to deliver one more PDF or HTML article to a desktop are almost so low as to consider them to be non-existent. Because the marginal costs are so small, a publisher can sell (or bundle) online access to institutions that have never owned a paper copy at prices that sound too good to be true.

Price discrimination only works when the producer has some monopoly power, which means that the customer cannot equally substitute one product or service for another. In an

exact sense, each journal is a monopoly since it is composed of a collection of unique articles not found in any other product — in essence, there is no substitutability. In a practical sense this is not completely true. Undergraduates will often require three articles on a particular topic, and so the substitutability for any article is very high, as long as it is still on topic.

There are three different *degrees* of price discrimination, and the distinctions between these are important for the publishing economy.

First Degree

First-degree price discrimination means that a producer (publisher) can sell every unit (journal article) at the maximum price that each consumer is willing to pay. In reality, a publisher currently lacks the information necessary to put this pricing model into practice. When you try to download an article from a journal to which your library does not subscribe, you (and everyone else) are pitched the same price.

In the case of journal subscriptions, the use of subscription vendors prohibited publishers from knowing too much about their customers. Publishers however, were not entirely in the dark. The construction of the ISI Citation Reports in the mid 1970's provided publishers with a good idea of how important their journals were to the academic community as a whole. These citations reports however did not allow publishers to find out how their journals were being used at *each institution*. In effect, they did not have enough information to charge differential prices to each institution based on what publishers believed libraries were willing to pay.

continued on page 26

Bradford's Distribution, the 80/20 Rule, and Patterns ... from page 22

these two areas tend to be interfiled in the ranked listings of journals, thus skewing the expected rate at which journal use drops.

In an attempt to reduce the effect of interdisciplinarity on the results, we extracted full-text usage statistics for education journal titles in **The College of Saint Rose** data set. Even within the education discipline, however, Bradford's distribution did not match the data. We found that "hits for 1" (again, we used the top three journals as "1") were 11,430. The Bradford multiplier was 5, such that hits were 12,195 for n , 48,018 for n^2 , and 37,713 for n^3 . (Hits for n^4 could not be calculated as there were not enough education titles in the set to do so.) Therefore, these results would suggest that interdisciplinarity does not account for Bradford's distribution not matching our data for college communities' use of interdisciplinary full-text databases.

Other possible reasons why Bradford's model does not fit this data include the fact that student use of online journals may differ from journal use by publishing professionals; changes in user behavior between

print journals and online sources; and characteristics of this particular sample. In addition, the reliability of online usage statistics remains to be determined.

References

- Bradford, Samuel C. (1953). *Documentation*, 2nd ed. London: C. Lockwood.
- Davis, Philip M. (2002). Patterns in Electronic Journal Usage: Challenging the Composition of Geographic Consortia, *College and Research Libraries*, 63, 484-499.
- Lockett, Mary W. (1989). The Bradford Distribution: A Review of the Literature, 1934-1987. *Library & Information Science Research*, 11, 21-36.
- Nisonger, Thomas E. (1998). *Management of Serials in Libraries*. Englewood, CO: Libraries Unlimited, Inc.
- Oluic-Vukovic, Vesna. (1997). Bradford's Distribution: From the Classical Bibliometric "Law" to the More General Stochastic Models. *Journal of the American Society for Information Science*, 48(9), 833-842.
- Trueswell, Richard L. (Jan. 1969). Some behavioral patterns of library users: the 80/20 rule. *Wilson Library Bulletin*, 43, 458-461.