Fame, Citation, and Collection Development

Howard D. White

Introduction

This paper draws conclusions from two lines of research I have pursued in recent years: the study of how authors cite and the study of how librarians develop collections in research libraries. In particular, I am concerned with whether citers are influenced by the *fame* of authors they cite and with whether librarians are influenced by the *fame* of the books they buy. Interestingly, my data on this question show that there are strong parallels in the behavior of citers and of librarians as they act over time. Both groups make nonrandom choices that result in similar bibliometric distributions. Bibliometrics is usually concerned with data that are aggregated at the level of journals, organizations, and scientific specialties — not with choices at the level of individual persons. But my data show that the latter kind of data yield highly interpretable patterns, and I will stress the parallel structure of these patterns in my paper.

In the case of citers choosing authors to cite, it is possible to imagine different strategies that follow contradictory advice:

(1) Cite works by *famous* authors because their names will reflect prestige on your claims and help persuade your readers that your arguments are worth heeding. (This advice might be given especially to young scholars and scientists at the beginning of their careers.)

(2) No, cite works by *little-known* authors because then everyone will be impressed with your scholarship in bringing these obscure works to light.

(3) No, cite authors at *any level of fame* if they are relevant to the argument you are making.

In the case of collection developers in research libraries, similar contradictory strategies could be advised:

(1) Buy mostly *famous* works because they have proved their popularity: those are the ones that everyone will want.

(2) No, buy *little-known* works because only those are worthy of the attention of true scholars. Famous works are not scholarly: they are only for the general public.

(3) No, buy works *at any level of fame* if you think they will be relevant to the needs of your users.

To see which of these alternatives is followed, we will look at the citing behavior of individual scholars and scientists, and the book-collecting behavior of individual librarians. We will find they exhibit the same patterns with regard to the fame of authors that citers cite and the fame of books that librarians buy.

Methods

I measure fame of cited authors by their citation counts in the databases of the Thomson Institute for Scientific Information — the well-known ISI citation indexes, Scisearch, Social Scisearch, and Arts & Humanities Search as supplied by Dialog. Using Dialog's RANK command, one can obtain the ranked frequencies with which an individual author has cited other authors in his or her papers over time. In White (2001, 2004), I have called this an author's *citation identity*. In the following example, we see the record of the British sociologist G. Nigel Gilbert in 26 articles. (Only the top 20 listings are given out of 357 in the total record.) Under "Items Ranked," we see that Gilbert has cited himself in 15 of the 26 articles. Others he has cited include Thomas S. Kuhn in 9 articles, Robert K. Merton in 5 articles, and so forth. But the *fame* data here are the counts under "Items in File." These show, for example, that Gilbert has himself been cited in 731 articles in the field of sociology — a good number, but nowhere near the 10,631 articles that cite Kuhn or the 9,087 articles that cite Merton. So the "Items in File" counts give us a reasonably objective metric for establishing a cited author's fame, and we can see how an individual citing author, here Gilbert, uses authors who are famous to varying degrees.

DIALOG RANK Results (Detailed Display)

RANK: S1/1-26 Field: CA= File(s): 7
(Rank fields found in 26 records -- 357 unique terms)

RANK No.	Items in File	Items Ranked	%Items Ranked	Term
1	731	15	57.7%	GILBERT GN
2	842	9	34.6%	MULKAY M
3	10631	8	30.8%	KUHN TS
4	388	8	30.8%	MULKAY MJ
5	1189	7	26.9%	COLLINS HM
6	3923	6	23.1%	LATOUR B
7	9087	5	19.2%	MERTON RK
8	1512	5	19.2%	POTTER J
9	1227	5	19.2%	UK OFF POP CENS S
10	1057	4	15.4%	BENDAVID J
11	249	4	15.4%	BRANNIGAN A
12	1579	4	15.4%	COLE S
13	1460	4	15.4%	CRANE D
14	379	4	15.4%	DALE A
15	295	4	15.4%	GASTON J
16	679	4	15.4%	HAGSTROM WO
17	1120	4	15.4%	LAW J
18	515	4	15.4%	MULLINS NC
19	464	4	15.4%	RAVETZ JR
20	891	4	15.4%	WOOLGAR S

The corresponding data for establishing the fame of a book is the number of libraries that hold it in OCLC's WorldCat, the international union catalog. Two examples from the literature in Library of Congress class DS902, which is history of Korea, may suffice to show holdings counts. The first book, in English, is held by 2,197 libraries, and so is taken to be a famous book in this literature. The second book, in Korean, is held by only 13 libraries, and so is taken here as a work little-known or obscure to most of the librarians WorldCat serves, who tend to be English speakers.

A handbook of Korea.
Publication: [Seoul, Korea] : Korean Overseas Information Service, 1993
Document: English : Book
Libraries Worldwide: 2197

한국의궁궐 /
Han 'guk u~i kunggwo~l /
Author: 이강근,Yi, Kang-gu~n.
Publication: 서울 : 대원사, 1991.; So~ul : Taewo~nsa, 1991
Document: Korean (Hide non-Roman characters) : Book
Libraries Worldwide: 13
More Like This: Search for versions with same title and author | Advanced options ...
See more details for locating this item

In the case of books, the person corresponding to a citing author, such as G. N. Gilbert, is one or more individual librarians who develop collections at research libraries. Although I do not know these persons by name, I can identify them by their libraries — for example, "Berkeley" for the collection developer at the University of California at Berkeley. (There may well be more than one person who has developed a library's collection in a subject such as Korean history, but there would not be many, and we can regard them as "an individual" carrying out a collection policy.) And just as we can place authors on a scale of fame and determine how individual citers cite them, so we can place books on a scale of fame and determine how individual librarians collect them.

In WorldCat, all the bibliographic records of books in a literature can be retrieved at once. WorldCat also has a module whereby the holdings counts of the entire literature can then be displayed in compact form on what I will call the WorldCat holdings-count scale. This special scale distributes the counts in bins that are not at all equal. Some intervals of the bins are units, some are tens, some are hundreds, and some are thousands. In this, the WorldCat holdings-count scale resembles a *logarithmic* scale rather than a linear one. The following example shows the literature of Korean history (Library of Congress class DS902) as distributed over the holdings-count scale, which is the scale of fame for books.

Search | Clear

2000-2499	1
1500-1999	2
1000-1499	3
900-999	1
800-899	3
700-799	3
600-699	1
500-599	4
400-499	9
300-399	10
200-299	24
150-199	12
100-149	34
75-99	21
50-74	31
25-49	86
10-24	206
5-9	229
2-4	403
1	351
0	4

Search	Clear
1500-1999	2
800-899	2
700-799	2
600-699	1
500-599	2
400-499	4
300-399	5
200-299	9
150-199	7
100-149	15
75-99	5
50-74	12
25-49	21
10-24	75
5-9	77
2-4	63
1	17

At the top, one book falls in the category of being held by from 2,000 to 2,499 libraries (this is *A Handbook of Korea*, as seen above). At the bottom, four books have been cataloged but not yet reported as being held by any library, and some 351 books are held by only one library each. Top and bottom show the extremes of fame.

This same module in WorldCat allows us to display not only the holdings counts for an entire literature but also for a *collection* of that literature on the same scale. A collection is retrieved by entering an OCLC code-name for a library and then retrieving part of the literature that is held by that library. Thus, we can see how individual collection developers behave with respect to buying along the scale of fame for books.

This is Berkeley's record of acquisitions in Korean history on that scale:

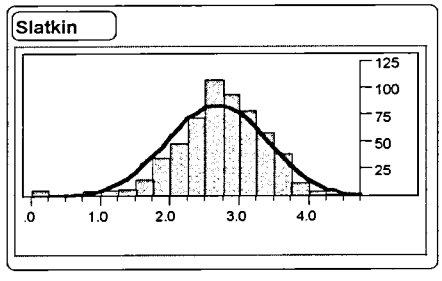
Results for Citers

In White (2004), I established that the scale of fame for cited authors, such as those in the G. N. Gilbert example above, is not linear but logarithmic. That is, if the raw citation counts under "Items in File" are converted to base-10 logarithms and placed in rank order, one finds that they have an approximately normal distribution over the following scale:

Log_{10} scale	*Label*	*Raw citation count*
0.01—1	Obscure	1—10
1.01—2	Recognized in specialty	11—100
2.01—3	Well known in discipline	101—1,000
3.01—4	Well known beyond discipline	1,001—10,000
4.01—5	World famous	10,001 and higher

The reason for using a log scale rather than a linear one is that gradations in fame of cited authors are not discernible from count to count (such as from 279 to 280 citations), but they *are* discernible across orders of magnitude of counts. For example, authors with only 100 citations, who score 2 on the log scale of fame, are discernibly less well known than authors with 5,000 citations, who score 4 on it. I currently have fewer than 600 citations in Social Scisearch, and I am surely less famous than Noam Chomsky, who has more than 10,000. I am a 3 ("well known in my discipline" according to the table); he is a 5 (world famous across many disciplines).

However, the real finding of White (2004) is that the distribution of every single citing author I studied, including 10 information scientists, four humanists, six natural scientists, and eight sociologists of science, was log-normal over the scale of fame. It did not matter whether they were young or old, male or female, recent PhDs or established masters, they all had distributions like the following one, which is for Montgomery Slatkin, a population geneticist at UC Berkeley. At bottom one sees the scale of fame, from very obscure authors at left to world-famous authors at right:

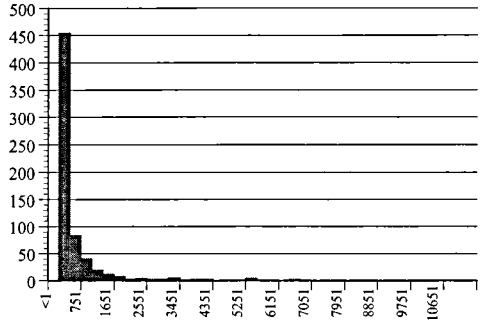


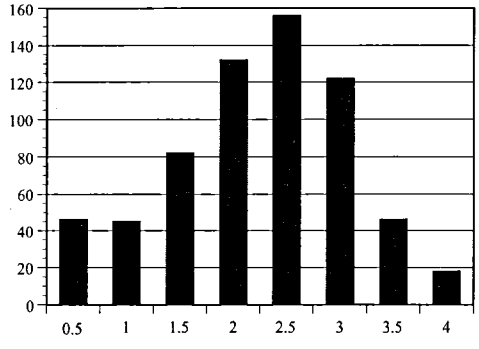
Slatkin cites relatively few famous authors (4's and above) and relatively few obscure authors (1's and below). Most of his citations go to authors who are middling in fame (2's and 3's — people known in their specialties or their disciplines but not beyond). By my logic, we can conclude that he is not trying to impress people by concentrating his citations on the very famous, nor is trying to impress them by referring only to obscure authors to show the depth of his scholarship. He does not *avoid* the famous or the obscure, but he obviously is not seeking them out. His pattern suggests that he is citing simply because other authors have written things he finds relevant to his own work, and he does not care where they stand on the scale of fame.

I have cited hundreds of other authors over the years, and I can display my own record

using the same techniques. The charts below, made with different software than the Slatkin chart, show two views. The first is a plot of the raw citation counts of authors I cite: it is the classic "reverse-J" curve of bibliometrics. The second is exactly the same data transformed into base-10 logarithms, with the same logarithmic scale of fame along the bottom axis. It is true that I have tended to cite somewhat more heavily at the "obscure end" than Slatkin; that may be because I have written lengthy review articles that covered works by relatively unknown writers. But clearly the distribution is roughly log-normal. Were I to reproduce the equivalent charts for G. N. Gilbert (or any other author), they would look similar. In summary:

1. Citers do *not* cite only famous authors.
2. They do *not* cite only obscure authors.
3. They cite across all levels of fame.
4. *Most* of the authors they cite are known within a discipline but are neither famous nor obscure.
5. Conclusion: They cite authors at any level of fame who are *relevant to their claims.*





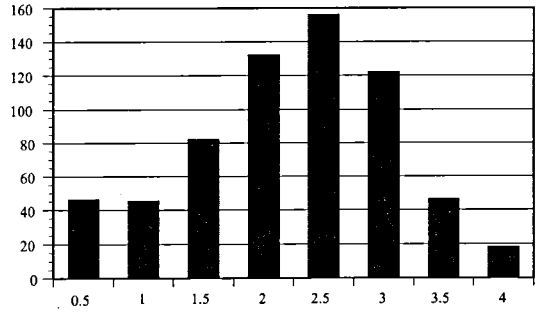
Results for Collection Developers

Turning to the analysis of how collection developers behave with respect to the fame of the books, I must first note that author citation counts run from 1 to more than 100,000. My logarithmic scale of fame for cited authors thus has intervals of 1 and runs from 1 through 5. However, in most literatures, holdings counts for books run from 1 to fewer

than 3,000. A logarithmic scale of fame for books is thus much more condensed.

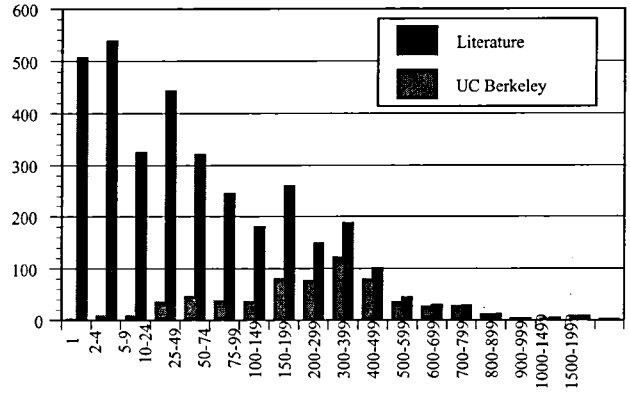
If every book held by fewer than 25 libraries is deemed *highly* obscure, the upper part of the WorldCat holdings-count scale can be approximated as 25, 50, 100, 20, 400, 800, 1600, and 3200. The base-10 log equivalents have equal intervals of .3; they run 1.4, 1.7, 2, 2.3, 2.6, 2.9, 3.2, and 3.5. I mention this because I will use the WorldCat scale, rather than a strict log scale, to display collection developers' acquisitions. (This is for convenience; holdings-count data for both literatures and collections can be quickly moved from WorldCat to spreadsheet and graphics packages if the WorldCat scale is used.) Furthermore, for simplicity's sake, I will equate levels of fame with four broad, crude categories. If a book is held by fewer than 150 libraries, I will call it "little-known." Books held by 151 to 400 libraries will be "known in specialty"; books held by 401 to 800 libraries, "known in discipline." Books held by more than 800 libraries (which are fairly rare) will be considered "famous" or library best-sellers.

With this background, let us look at UC Berkeley's collection in HM51, which is the Library of Congress class for general theory in sociology:



Again, we see an approximately log-normal distribution — one that implies Berkeley is not acquiring only famous books or only obscure books. Indeed, it is books in the *middle range* of fame that account for the most purchases in absolute numbers. These are books known within a specialty or a discipline, but neither famous nor obscure.

If one simultaneously plots the *literature* of HM51 with Berkeley's *collection* in HM51, one sees further that Berkeley has bought just about everything there is to buy at the top end of the scale; it has acquired almost all of the library best-sellers. At the bottom end of the scale, it has bought relatively few of the books that are library worst-sellers. It gets dozens of them, but these are "different dozens" from what other libraries are buying.



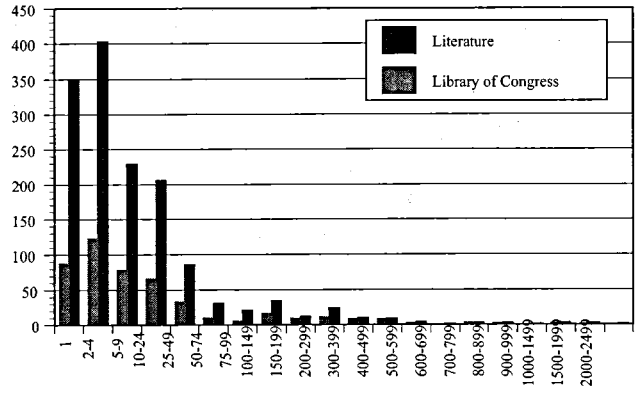
So again we see a policy of even-handedness across the scale of fame. The most popular books are indeed bought. At the other end of the scale, some highly specialized items are acquired, but not big proportions of what there is to buy. The great bulk of what is bought are books that some researchers in a specialty or discipline might recognize, but nothing more than that. This pattern occurs constantly in collection development in research libraries like Berkeley's: the relation between literature and collection seen in the chart is highly typical whenever big collections that support serious researchers are being developed.

My summary exactly parallels what I said of citers before:

1. Librarians in research libraries do *not* buy only famous books in a literature.
2. They do *not* buy only little-known books in a literature.
3. They buy across all levels of fame.
4. *Most* of the books they buy are known within a discipline but are neither famous nor little-known.
5. Conclusion: They buy books at any level of fame that are *relevant to the needs of their users.*

There is, however, an interesting exception to the roughly symmetric log-normal distribution of library collection development. It occurs when large numbers of books *in languages other than English* are considered necessary in a research specialty. When that happens, the distribution of the collection is skewed toward the low end of the scale of

fame. The Korean history collection of America's Library of Congress illustrates:



Here we see that the greatest number of acquisitions at the Library of Congress are books held by fewer than 10 libraries. The reason, of course, is that many of these books are in Korean or other Asian languages, which makes them unknown to all but a few American librarians. But the Library of Congress still gets them because they may be relevant to scholars who read Korean.

Speaking more generally, are citers and librarians aware of these strong patterns in their behavior? I think not. They are not consciously trying to *produce* these distributions on the scale of fame: they are simply behaving naturally over time. My cautious conclusion is that the distributions are produced by repeated choices of individual persons in specific situations. The choices are not random: rather, they can best be explained by a personal sense of relevance, and the distributions accumulate from that.

This idea should be explored further, because relevance is often said to be the key concept in information science.

REFERENCES

White, Howard D. 2001. Author-centered bibliometrics through CAMEOs: Characterizations automatically made and edited online. *Scientometrics*, 51: 607-637.

White, Howard D. 2004. Reward, persuasion, and the Sokal Hoax: A study in citation identities. *Scientometrics*, 60: 93-120.