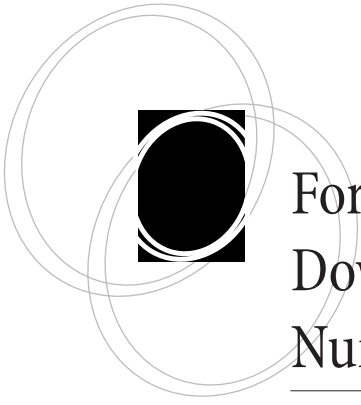
Project MUSE®
http://muse.jhu.edu



For Electronic Journals, Total Downloads Can Predict Number of Users

Philip M. Davis

abstract: Results from two multiple regression models involving HighWire journal subscriptions in 2003 for 16 participating universities in the United States, United Kingdom, and Sweden indicate a highly predictive relationship between the number of article downloads and the number of users—meaning that the size of a user population can be estimated by just knowing the total use of a journal. The relationship is consistent over time and across institutions and appears to be unrelated to the subject, size, or popularity of a journal. It is not consistent across publishers, however, suggesting that an "interface effect" may exist. The development of a Project COUNTER standard to deal with extreme or abnormal journal usage is necessary if we wish to compare the performance of journals across publishers.

Introduction

Most electronic journal usage reports inform librarians only of the frequency of article downloads during a period of time. They do not tell the librarian *who* was responsible for the downloads, *why* the articles were requested, or *how many* individuals were responsible for the downloads. The only definitive statement that a librarian can make about a journal that received 100 downloads last year was that it simply received 100 downloads. Was one individual responsible for downloading all 100 articles, or were there 100 individuals downloading one article apiece? Would it make a difference for collection development if we knew the size of the user population?

While many of the publishers of science, technology, and medicine titles now conform to standards established by the international Project COUNTER, the first level of compliance does not require the publisher to disclose the number of users.[1] Being able to estimate the user community based on the total number of downloads would be a

portal: Libraries and the Academy, Vol. 4, No. 3 (2004), pp. 379–392.
Copyright © 2004 by The Johns Hopkins University Press, Baltimore, MD 21218.

useful tool for librarians evaluating the utility and impact their journal subscriptions have in their communities.

This study examines whether the number of fulltext downloads can accurately predict (or estimate) the size of a user community. Specifically, it will test whether this relationship varies over time or by institution.

Methods

Monthly reports for Cornell University were downloaded in tab-delimited format from HighWire Press. Reports from other institutions were acquired by personal contact with electronic journal administrators and by posting a request for participation on an electronic listserv. Nineteen libraries from the United States, the United Kingdom, and Sweden responded with usage reports. Three institutions were eliminated from the analysis, because they employed technology that distorted their journal reports. The 16 participating institutions are listed in Table 1.

The 16 participating institutions represent a geographically distributed and relatively heterogeneous sample of medium and large institutions. In order to extract meaningful patterns from the data, the participating institutions needed to subscribe to a substantial number of HighWire journals and show significant use of at least some of the titles. The results of this study should be generalizable to all institutions; however, smaller institutions repeating this analysis may find that their results may vary. This is due to the inherent problems of working with smaller and more variable datasets.

The two primary variables under investigation in this study are the number of *fulltext downloads* and number of *unique IP addresses*. The number of unique IP addresses was used as a surrogate for the size of the readership community or number of "users." In general, it is impossible to measure directly the number of users without requiring that each individual log into a publisher's system. In an environment where the library is expected to maintain the confidentiality of the patron, IP address data are likely to provide the best estimate of the number of users. Whereas most IP addresses correspond to a single user, some IP addresses correspond to public computers in labs or libraries or to the library proxy server. Philip Davis and Leah Solla recently demonstrated in their study of e-journal use by IP addresses that computers that aggregate users may be regarded as a statistical sample of the entire user population.[2]

HTML and PDF downloads were combined for each journal to create the *fulltext* variable. The rationale for doing this was twofold: 1) most publishers combine HTML and PDF downloads to comply with Project COUNTER's Journal Report Level 1; and 2) based on pre-analysis that went into variable selection, combining HTML and PDF provides a more accurate prediction for number of users than either of these formats alone.

Multiple regression was the technique employed in this study. As an extension of simple regression, multiple regression allows the researcher to estimate the effect of many contributing variables at the same time. It has been used extensively in the field of medicine in order to estimate the effect of a drug, while controlling for other possible contributing factors such as age, gender, weight, and smoking, among others. In the field of library and information science, it has been used to estimate the effect of journal size, circulation, publisher, country, and other variables on the price of academic jour-

Table 1

Participating Institutions and number of titles included in dataset

Participating Institutions	HighWire Titles in dataset
Cal Tech	92
Cambridge University (UK)	66
Columbia University	90
Cornell University	150
Iowa State University	98
Kings College, London (UK)	122
Lund University (Sweden)	90
MIT	93
Rutgers University	93
University Georgia	110
University of Bristol (UK)	132
University of California, Berkeley	72
University of Liverpool (UK)	135
University of Nevada, Reno	78
University of North Carolina, Charlotte	46
Yale University	177

Notes: Because HighWire functions as an online host for publishers, the set of subscribed titles may not be consistent across all libraries. In addition, HighWire allows multiple individuals at the same institution to serve as journal administrators. For this reason, the data for certain institutions may not include every HighWire journal available to its patrons.

nals.[3] It was recently employed to evaluate the effect of instruction on the number of reference questions in ARL libraries.[4]

Like all statistical tests that depend on an analysis of variance, regression requires the dataset to adhere to four basic assumptions: 1) independence of observations, 2) normality in the distribution of the data, 3) constant variance along the relationship, and 4) for the purposes of linear regression, linearity among the relationships.[5] Usage data, like many other distributions in library and information science, are not normally distributed but severely skewed.[6] That is, the highest performing journals generally receive many times more downloads than the lowest performing ones. Without normalizing the data, these large journals would overly bias and obscure the results of the test.

To achieve normality and linearity, the dataset required transformation. The number of fulltext downloads and the number of IP addresses were expressed in the natural logarithm *ln* (pronounced "lawn") form. This transformation, however, poses some problems when dealing with a dataset that contains zeros, since it is logically impos-

sible to take the log of zero. To adjust for this problem, one download is added to each journal before log transformation. All continuous variables reported in this paper have been transformed using the *ln*+1 technique. Similar transformation of a skewed dataset containing a large number of zeros was used to analyze network traffic.[7] The data were gathered and processed in Excel and then exported into SPSS for analysis.

Data Models

Two regression models are used to estimate (or predict) the number of unique users across all journal titles. The first model controls for the effect of time (in months) on this relationship. The second one controls for the effect of different institutions on the model.

Model 1: Testing the Effect of Time

ln Unique IP addresses = B_0 (intercept) +
B_1 *ln* downloads +
$B_{2 \ldots 12}$ month (Jan + . . . Nov) +
$B_1 * B_2 + \ldots B_1 * B_{12}$ (Interaction between downloads and month) +
error

Model 2: Testing the Effect of Institution

ln Unique IP addresses = B_0 (intercept) +
B_1 *ln* downloads +
$B_{2 \ldots 15}$ institution (Cornell + Cal Tech + . . .) +
$B_1 * B_2 + \ldots B_1 * B_{15}$ (Interaction between downloads and institution) +
error

In both models, *month* and *institution* are coded as indicator (or *dummy*) variables. This allowed the researcher to test whether individual months or institutions have an effect, as well as their possible interaction with other variables. An interaction effect would indicate a change in slope of the regression line.

Model 1 is based on data from all 12 months reported for Cornell University. Model 2 is based on data for all participating institutions for the month of October 2003. If there is no significant monthly effect, then using a single sample month would be sufficient to test the effect of each institution. After running the regression models, residual error plots were generated and examined to determine whether the assumptions of linearity and equal variance could be met.

Observations

In general, *month* and *institution* had little (if any) effect on the number of *ln fulltext downloads*. Small observed effects could be explained by normal variation, missing values, or outlying data points.

Testing the Effect of Month (Model 1)

Scatter plots revealed a linear and very strong relationship between the number of *ln* downloads and the number of *ln* unique IP addresses. Figure A illustrates this relationship for all Cornell subscribed HighWire journals for 2003 (all months totaled). The regression coefficient, a measurement of the goodness of fit of the regression line, was 0.96—meaning that 96 percent of the variation in the model can be explained by just knowing the number of downloads.

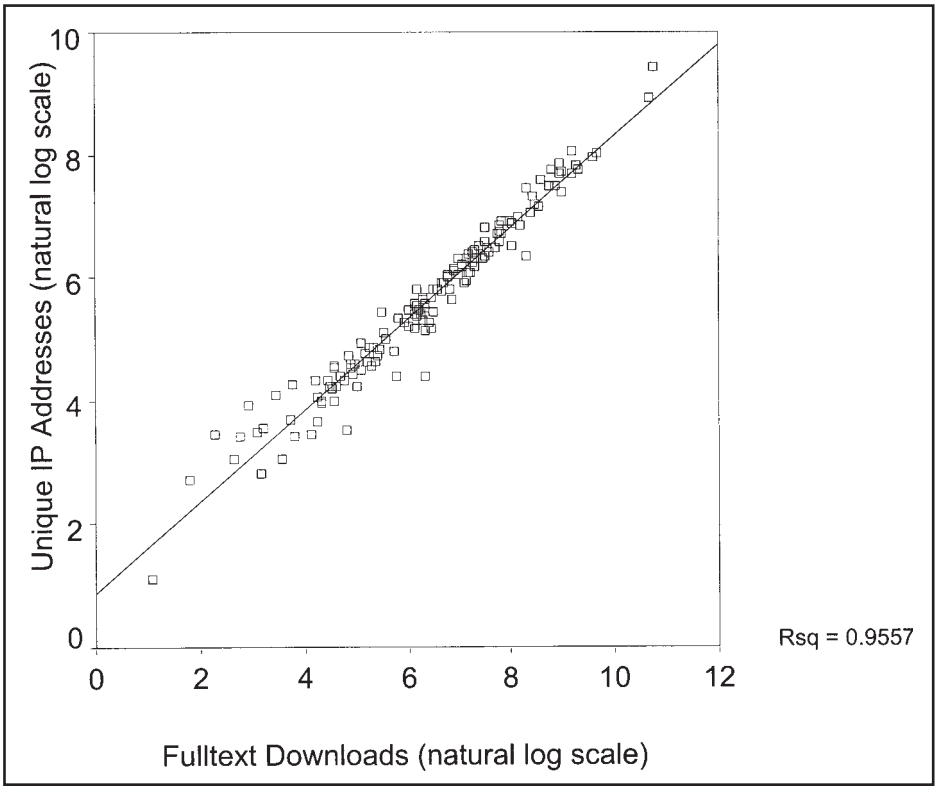


Figure A. Number of fulltext downloads predict the number of unique IP addresses ('users'), Cornell University 2003 (all months). Each datapoint represents a journal.

Table 2 (Appendix) lists the regression statistics for Model 1. The number of *ln* fulltext downloads is a very good predictor of the number of *ln* unique IP addresses. The regression coefficient for downloads was 0.71, and the confidence interval (C.I.) associated with this parameter ranged from 0.68 to 0.74. What this means is that for every additional *ln download*, the number of *ln IP addresses* is expected to increase by 0.71, and we are 95 percent confident that this number is somewhere between 0.68 and 0.74. Because both of these variables were log transformed, interpreting these results in raw downloads and raw users is a bit tricky and requires back-transformation. For ease of lookup, figure B was created, which illustrates a scaling effect present in the data. As

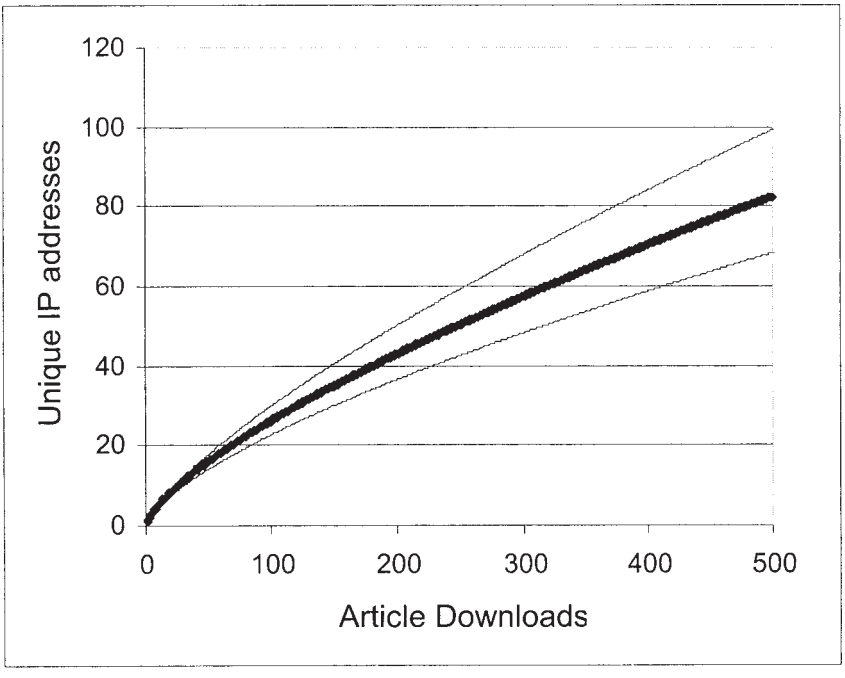


Figure B. Solution of regression equation back-transformed to illustrate scaling effect. Bold line represents predicted value, thin lines represent confidence intervals. The slope of the solution decreases as the number of downloads increase. The confidence intervals also increase in proportion to downloads.

journals receive more downloads, the proportion of additional users starts to decline. At the same time, the confidence interval associated with the size of the user population increases.

All monthly effects are compared to December, which served as the reference month. Since December was the highest use month for journals in 2003, all other monthly betas (B) were negative. As mentioned earlier, the interaction effect (i.e., *month * fulltext downloads*) indicates that there is a change in slope in the regression line. With the exception of September, no significant monthly interaction effects were detected. Post hoc analysis revealed that September illustrated very unusual behavior compared to other months at Cornell, with a larger number of journals registering zero use (figure C). It is likely that the beginning of the fall semester and a renewed emphasis on teaching may be responsible for this unusual behavior.

While the set of monthly variables had a significant effect on the model, it was very small. Adding the month variable to the regression model explained only an additional 0.8% of the variation. In a practical sense, it may be considered to have little (if any) effect.

Testing the Effect of Institution (Model 2)

Table 3 (Appendix) lists the regression statistics for Model 2. All institutional effects are compared to Yale, which served as the reference institution. The institutional variables

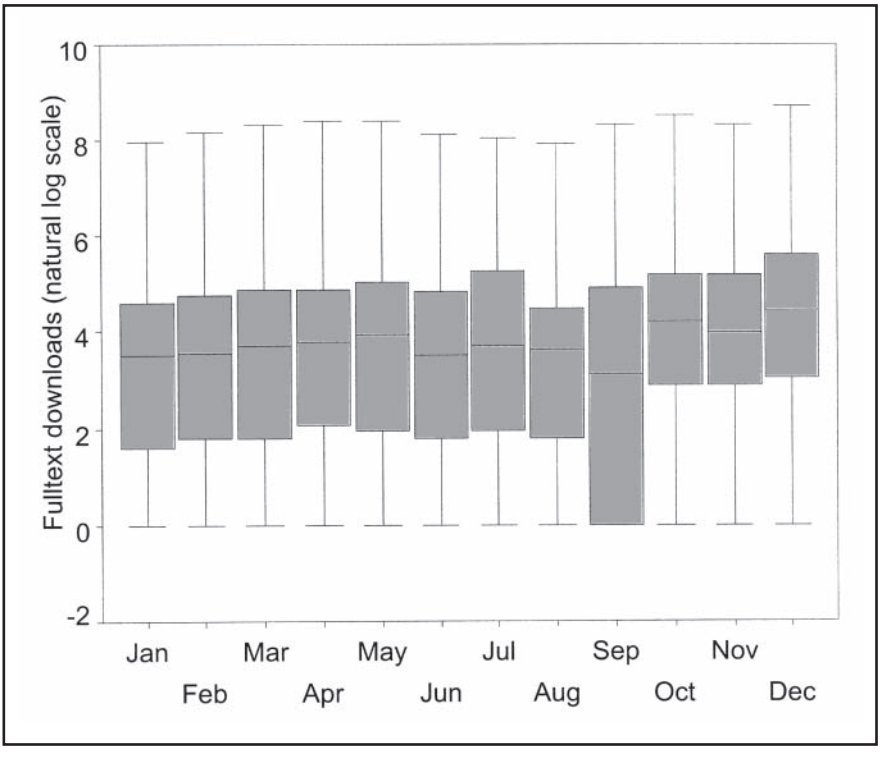


Figure C. Distribution of Journal Use by Month, Cornell University, 2003.

provide very little additional explanation in the regression model; in fact, knowing the institution explained only 0.7 percent additional variance. Post hoc analysis demonstrated that the University of Nevada might have demonstrated unusual behavior compared to other institutions. Figure D illustrates regression lines for each of the 16 participating institutions. Nevada's datapoints are highlighted with solid triangles, and its dashed regression line clearly deviates from the rest of the institutions. In comparison to datasets provided from other libraries, Nevada's use of HighWire journals was much lower. In addition, their dataset did not include several titles that are normally associated with high usage, such as *Proceedings of the National Academy of Sciences* (*PNAS*) and the *Journal of Biological Chemistry*. The absence of these titles from the upper right corner of the graph may have allowed for a deviation of their regression line.

The coefficient for fulltext downloads was 0.71 (C.I. 0.68 – 0.73), exactly the same number that we got from Model 1. The ratio of *ln* downloads to *ln* "users" appears constant irrespective of the size and location of the institution.

Discussion

Results from these two models suggest a consistent relationship between the log of fulltext downloads and the log of IP addresses across time and across institutions. In addition, this relationship appears to have high predictive value across an entire publisher's suite of journals irrespective of subject, size, or popularity.

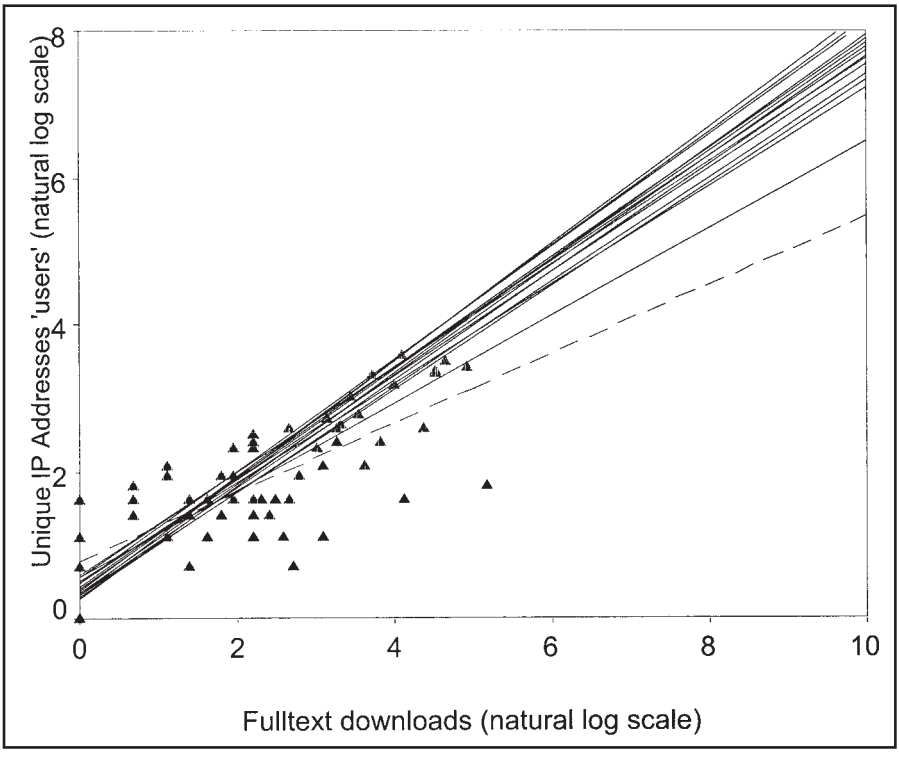


Figure D. Number of downloads predict number of unique IP addresses ('users') for all institutions. Each regression line represents an institution. One institution exhibited outlying tendencies. Its regression line is dashed and datapoints are highlighted by solid triangles.

Consistency Across Publishers

These results confirm the strong linear relationship between the number of log fulltext downloads and the number of log users as first reported by Davis and Solla in their study of 29 American Chemical Society (ACS) e-journals.[8] What was not consistent between the publishers was the slope of this relationship. The ratio of downloads to users for HighWire titles was 0.71 or 1.4 ± 0.3 *ln* downloads per *ln* user. For ACS titles the ratio was 0.093 or about 11 ± 3.5 *ln* downloads per *ln* user. In real (non-transformed) numbers, a journal that registered 100 downloads in a single month would represent 26 HighWire users but only 1.5 ACS users. Two possible explanations may account for this relatively large difference: 1) interface effect and 2) extreme use effect.

Interface Effect

The internal consistency of all journals in a publisher package suggest that this difference cannot be attributed to subject scope or size of user community. Despite the name, the ACS publishes journals in the field of food science, biochemistry, engineering, and physics. And while HighWire focuses on the biomedical literature, they also publish journals in general science, physical and social sciences. The differences in subject scope are unlikely to provide an explanation for the observed differences.

We may be dealing with the effect of the publisher interface to encourage (or dissuade) additional use through browsing and linking to other titles. Unlike the ACS, HighWire is an online provider, providing journal-hosting services. Most journals are found on their own separate Web sites and function essentially as independent journals from the perspective of the user. Not every journal takes advantage of all HighWire services, which include advanced linking using CrossRef, Medline, and the ISI's Web of Science. ACS journals are all found on the same server and include the same suite of linking services, including CrossRef, Medline, and a proprietary linking protocol called ChemPort. There are many tools used and pathways taken by patrons discovering online journals. The author recently published a study of how users are referred to ACS journals.[9]

In a strict sense, it is impossible to directly estimate an interface effect when comparing different content on different sites. Such an experiment would require the same journals to be hosted by different providers so that the effect of an interface could not be confused with content. Consequently, estimating the interface effect can only be done by controlling for all other possible contributing factors.

Extreme Use (or Abuse) Effect

For journal systems, usage is not evenly distributed across all users—a small minority of heavy users is responsible for the majority of article downloads. These few individuals can greatly affect the findings of any usage study. In the well-documented British SuperJournal project[10] researchers grouped users into four categories based on their frequency of use—from the "enthusiastic user" who downloads a large number of articles from a large number of journals to the "restricted user" who downloads a few articles from a small number of journals.[11] In a transaction log analysis of Elsevier journals available from ScienceDirect, Taiwanese researchers illustrated that the same skew of use was concentrated on a small number of IP addresses and speculated that these extreme users may be attributed to proxy servers, firewalls, or other servers that aggregate users under a single IP address.[12] In their analysis of ACS journal use by IP address, Davis and Solla tracked down the origin of extreme users. In all cases, with the exception of the library proxy server, extreme downloads were associated with individually owned computers.[13] Computers that aggregate users, such as a shared computer in a library or computer lab, do not appear to have much (if any) effect on the general relationship between downloads and users.

For journal systems, usage is not evenly distributed across all users—a small minority of heavy users is responsible for the majority of article downloads. These few individuals can greatly affect the findings of any usage study.

Whereas the majority of Cornell users download one to two articles from ACS journals per month, approximately 3 percent per month of Cornell IPs are associated with downloading more than 100 articles. In some cases, single users have downloaded several thousand articles in a single month. The inclusion of these extreme or "enthusias-

tic" users may at least partially explain the higher average number of articles per user for ACS over the HighWire results since they would have the effect of skewing the results.

Personal correspondence with both HighWire Press and ACS Press identified that both publishers incorporate systems that block potential abuse from their online systems.[14] The methods they use, however, are different; and this may at least partially explain the difference in regression slopes between the publishers.

Implications for Project COUNTER

Project COUNTER was formed in 2002 by publishers, vendors, and librarians to develop an "internationally accepted, extendible Code of Practice . . . that will enable vendors to provide these requested statistics in a way which all parties can trust to be consistent, credible and compatible."[15] Many of the largest publishers of STM journals have become compliant members. Their most notable achievement to date has been the creation of a standard to count e-journal article requests. Based on the pioneering work by Marthyn Borghuis of Elsevier, the standard prevents double counting of the same article within a fixed period of time.[16] Multiple requests for the same article within 30 seconds are disregarded for PDF format versions and within 10 seconds for HTML. There is no current standard for dealing with users that demonstrate extreme or abusive behavior. Without standards for dealing with extreme users, one primary goal of Project COUNTER—to enable librarians to be able to compare usage of journals across different publishers—may not yet be realizable.

Implications for Librarians

The size of user communities is an important piece of information for librarians involved in collection development. It is especially important when the librarian is responsible for making journal cancellations and wants to minimize the impact of a cancellation on his or her community. While usage statistics cannot tell us *who* used a resource or *why* articles were downloaded, they can inform us *how many* individuals were involved. Being able to look at raw usage data and estimate the size of a user community may be a very powerful tool in the decision-making process, especially if a universal relationship exists between the number of downloads and the number of users that can be applied across all journals.

The size of user communities is an important piece of information for librarians involved in collection development. It is especially important when the librarian is responsible for making journal cancellations and wants to minimize the impact of a cancellation on his or her community.

The inclusion of extreme users has the effect of artificially distorting total use and projecting a higher utility for the user community. In a practical sense, a publisher that counts abnormal use has a much

lower cost per use ratio and may escape the eye of a discriminating librarian when journal cancellations are required.

Finally, it should be noted that research done at Stanford University on 14 HighWire journals indicated that many users are downloading the same article first in HTML format for the purposes of browsing and then in PDF for the purpose of printing.[17] Librarians should not confuse number of downloads with the number of unique downloads.

In summary, there appears to be a strong and linear relationship between the number of log downloads and number of log unique IP addresses that is constant across an entire suite of publisher titles irrespective of journal size, subject, or popularity. The ability to accurately predict the number of "users" based merely on the total number of downloads has been demonstrated to be consistent across time and across institutions. Because the model requires that the variables be log transformed prior to analysis, a back transformation is required to retrieve raw output. At present, there appears to exist a very large "interface effect" that may make it impossible to compare journals across publishers. The inclusion of a Project COUNTER standard to deal with extreme or abnormal journal downloads is necessary if we are to move closer to this goal.

Acknowledgements

The author wishes to acknowledge Bill Walters at St. Lawrence University, Terry Bucknell at the University of Liverpool, and Bob Molyneux at NCLIS for their review of this manuscript before submission.

Philip M. Davis is life sciences bibliographer, Albert R. Mann Library, Cornell University, Ithaca, NY; he may be contacted via e-mail at: pmd8@cornell.edu.

Appendix

Table 2

Regression Statistics for Model 1, testing the effect of time

Dependent Variable: Unique IPs ('users')
N=1708; Rsq Adj = 0.942

Parameter	B	Sig. (P value)	95% Confidence Interval Lower	Upper
Intercept (B_o)	0.71	0.000	0.56	0.86
Jan	-0.56	0.000	-0.74	-0.37
Feb	-0.49	0.000	-0.67	-0.30
Mar	-0.44	0.000	-0.62	-0.25

Table 2. cont.

Parameter	B	Sig. (P value)	95% Confidence Interval Lower	Upper
Apr	-0.24	0.012	-0.42	-0.05
May	-0.17	0.073	-0.36	0.02
Jun	-0.42	0.000	-0.61	-0.24
Jul	-0.32	0.001	-0.50	-0.13
Aug	-0.28	0.003	-0.47	-0.10
Sep	-0.63	0.000	-0.80	-0.46
Oct	-0.05	0.627	-0.26	0.16
Nov	-0.09	0.393	-0.31	0.12
Dec ‡	0	.	.	.
Fulltext downloads (FT)	**0.71**	**0.000**	**0.68**	**0.74**
Jan * FT	0.04	0.102	-0.01	0.08
Feb * FT	0.04	0.093	-0.01	0.08
Mar * FT	0.02	0.366	-0.02	0.06
Apr * FT	-0.01	0.775	-0.05	0.04
May * FT	-0.06	0.011	-0.10	-0.01
Jun * FT	0.00	0.826	-0.04	0.05
Jul * FT	-0.03	0.192	-0.07	0.01
Aug * FT	0.00	0.938	-0.04	0.05
Sep * FT	0.09	**0.000**	0.05	0.13
Oct * FT	-0.02	0.477	-0.06	0.03
Nov * FT	-0.01	0.746	-0.06	0.04
Dec * FT ‡	0	.	.	.

‡This parameter is set to zero because it is redundant.

Table 3

Regression Statistics for Model 2, testing the effect of institution

Dependent Variable: Unique IPs ('users')
N=3472; Rsq Adj = 0.941

Parameter	B	Sig. (P value)	95% Confidence Interval Lower	Upper
Intercept (B_o)	0.48	0.000	0.36	0.60
Bristol	0.02	0.845	-0.18	0.22
CalTech	-0.09	0.331	-0.28	0.10

Parameter	B	Sig. (P value)	95% Confidence Interval Lower	Upper
Cambridge	-0.28	0.021	-0.52	-0.04
Columbia	-0.14	0.170	-0.34	0.06
Cornell	0.08	0.353	-0.09	0.26
Iowa	-0.07	0.468	-0.26	0.12
Kings College	-0.15	0.211	-0.38	0.08
Liverpool	-0.16	0.064	-0.34	0.01
Lund	-0.28	0.036	-0.54	-0.02
MIT	-0.08	0.496	-0.31	0.15
Nevada	0.27	0.003	0.09	0.44
Rutgers	0.09	0.450	-0.14	0.31
Berkeley	-0.20	0.188	-0.51	0.10
UGA	-0.13	0.175	-0.32	0.06
N Carolina	0.07	0.513	-0.14	0.28
Yale ‡	0	.	.	.
Fulltext downloads (FT)	0.71	0.000	0.68	0.73
Bristol * FT	-0.03	0.168	-0.08	0.01
CalTech * FT	0.07	0.002	0.03	0.12
Cambridge * FT	0.05	0.046	0.00	0.10
Columbia * FT	0.06	0.010	0.01	0.11
Cornell * FT	0.01	0.592	-0.03	0.05
Iowa * FT	0.04	0.105	-0.01	0.09
Kings College * FT	0.09	0.000	0.04	0.14
Liverpool * FT	0.00	0.828	-0.05	0.04
Lund * FT	0.03	0.362	-0.03	0.08
MIT* FT	0.07	0.003	0.03	0.12
Nevada * FT	-0.21	0.000	-0.27	-0.15
Rutgers * FT	0.01	0.613	-0.04	0.06
Berkeley * FT	0.06	0.039	0.00	0.12
UGA * FT	0.04	0.128	-0.01	0.09
N Carolina * FT	-0.09	0.019	-0.16	-0.01
Yale * FT ‡	0	.	.	.

‡This parameter is set to zero because it is redundant.

Notes

1. Project COUNTER, "COUNTER: Counting Online Usage of Networked Electronic Resources," http://www.projectcounter.org/index.html (accessed 4/21/04).
2. Philip M. Davis and Leah Solla, "An IP-Level Analysis of Usage Statistics for Electronic Journals in Chemistry: Making Inferences About User Behavior," *Journal of the American Society for Information Science and Technology* 54, 11 (September 2003):1062–8.

3. George A. Chressanthis and Jane D. Chressanthis, "The Determinants of Library Subscription Prices of the Top-Ranked Economics Journals: An Economic Analysis," *Journal of Economic Education* 25, 4 (Fall 1994): 367–82; ———, "A General Econometric Model of the Determinants of Library Subscription Prices of Scholarly Journals: The Role of Exchange-Rate Risk and Other Factors," *Library Quarterly* 64, 3 (July 1994): 270–93; H. Craig Petersen, "University Libraries and Pricing Practices by Publishers of Scholarly Journals," *Research in Higher Education* 31, 4 (August 1990): 307–14; and———, "The Economics of Economics Journals: A Statistical Analysis of Pricing Practices by Publishers," *College & Research Libraries* 53, 2 (March 1992): 176–81.
4. E. Stewart Saunders, "The Effect of Bibliographic Instruction on the Demand for Reference Services," *portal: Libraries and the Academy* 3, 1 (January 2003): 35–9.
5. William Dale Berry, *Understanding Regression Assumptions* (Newbury Park, CA: Sage, 1993).
6. Stephen J. Bensman, "Bibliometric Laws and Library Usage as Social Phenomena," *Library Research* 4, 3 (Fall 1982): 279–312.
7. Melvin J. Hinich and Robert E. Molyneux, "Predicting Information Flows in Network Traffic," *Journal of the American Society for Information Science and Technology* 54, 2 (January 15, 2002): 161–8.
8. Davis and Solla.
9. Philip M. Davis, "Information Seeking Behavior of Chemists: A Transaction Log Analysis of Referral URLs," *Journal of the American Society for Information Science and Technology* 55, 4 (February 15, 2004): 326–32.
10. David J. Pullinger, *The SuperJournal Project*, *British Library R&D Report 6126* (Philadelphia: Institute of Physics Publishing, 1994).
11. Ken Eason, Sue Richardson, and Liangzhi Yu, "Patterns of Use of Electronic Journals," *Journal of Documentation* 56, 5 (September 2000): 477–504.
12. Hao-Ren Ke et al., "Exploring Behavior of E-Journal Users in Science and Technology: Transaction Log Analysis of Elsevier's ScienceDirect OnSite in Taiwan," *Library & Information Science Research* 24, 3 (2002): 265–91.
13. Davis and Solla.
14. John Sack, HighWire Press, personal e-mail January 30, 2004; Dean Smith, American Chemical Society Publications, personal e-mail January 30, 2004.
15. Project COUNTER.
16. Marthyn G.M. Borghuis, *What to Count and What Not? A White Paper on the Filters to be Applied to a Web-Server Log File Before Usage-Analysis and Reporting Can Start* (Amsterdam: Elsevier ScienceDirect, 2000).
17. Institute for the Future, *E-Journal User Study: Report of Web Log Data Mining* (Menlo Park, CA: Stanford University: 2002).