A Study of Rank Distributions of Journals and Articles

I.K. Ravichandra Rao and Bibhuti Bhusan Sahoo

ikrrao@hotmail.com, bibhuti@isibang.ac.in
Documentation Research and Training Centre, Indian Statistical Institute, 8th Mile Mysore Road, Bangalore 560059, India

Abstract
Two bibliographies (one in mathematics and the other one in Physics) are analyzed to study the relation among the journals, references and the citations as well as both the rank and size distributions of citations. It has been observed that the number of citations (z) received by a journal can be estimated using a log model i.e. z = ay – by log (y) (Basu's Model); y is the number of references in a subject in x most journals. Further, it has been observed that Basu's model with two free parameters (a and b), y = ax - bx log (x), fits very well to the observed data on the rank distribution of articles, where y is the citations received by the x most productive articles. It has also been observed that the size distribution of citations follows a negative binomial distribution, implying that the distribution of citations in a bibliography is a manifestation of success-breeds-success phenomenon. The Spearman rank correlation coefficient for the data on rank distributions of journals (based on articles and citations) is only 0.0972515, indicating that the ranks are quite different from each other; however, among the top 10 journals, 9 journals are common in both the ranked list. However, this is not true for the group consisting of least productive journals.

Introduction

Bradford (1934) observed that 'if scientific journals are arranged in order of decreasing productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus, when the zones will be 1: n^2 /n^3… ." Since then many have worked in this area and observed that Bradford's law holds well for a "complete bibliography of a well defined subject. In some cases, however, a "Groos droop" may be seen in the semi-logarithmic curve, indicating that the bibliography is incomplete and the subject is of inter-disciplinary in nature. Leimkulher derived a most widely accepted model of Bradford's law (1967). Brookes (1969) argued that the core and linear part of the semi-logarithmic curve (the Bradford curve) needed to be expressed as two separate equations. His model however does not explain the "droop." Vickery (1948) has pointed out the ambiguity" between the verbal and graphical statement of Bradford's law. Egghe (1990) derived a generalized Leimkulher's model, which fits even the Groos droop. Basu (1992) derived a probabilistic model of the most probable unequal distribution of a set of articles in a set of journals, called the Random hierarchical model. Her model uses an equation for the entire Bradford curve including the Groos droop. Almost all these models are empirical in nature and there is a need for re-examining the models in terms of:

- Principles underlying their derivations;
- Fitting the model to the observed data (to determine how far it is good)

Objectives Of The Study

The purpose of this study is however not to revisit Bradford's law or similar other laws. But to re-examine how far this model can be applied or extended to study (or to predict):

- The number of citations received by a set of articles (say, y) which are published in top most x journals on a given subject.
- The rank distribution of articles in a bibliography of a given subject, based on the citations received.
- The size distribution of citations in a bibliography.

Proceedings of ISSI 2005, pages 140–148.
Edited by P. Ingwersen and B. Larsen.

An attempt is also made to study whether or not there is any difference between the two ranked lists of journals – one list based on the number of articles it contains and the other list is based on the number of citations received by the relevant articles.

Data Collection

For the purpose of this study, a "bibliography for automorphic and modular forms, L-functions, and representation theory" [1] was analyzed and the following data were collected, after arranging the journals in decreasing productivity (the productivity is measured in terms of the number of articles a journal contains):

- Partial sum of journals (x) (Ref. Col. (b) in Table 1)
- Partial sum of articles (y), contained in x journals (Ref. Col. (d) in Table 1)
- Partial sum of citations received (z) by the y articles in x journals (Ref. Col. (e) in Table 1)

There were 151 references (articles) in the bibliography; these are covered in 43 journals, in mathematics. For each of the references, the number of citations received (z) (from 1993 to 2000) were obtained using SCI CDs. The data is given in Table 1 (Ref. Col. (f)). Each of the references is also arranged in order of decreasing productivity (the productivity is measured in terms of citations received), in order to study the rank distribution of articles. This data is given in Table 2.

Table 1: A Distribution of Journals and References (Mathematics)

No. of Journals	**Partial sum of Journals (x)**	**No. of References**	**Partial sum of References (y)**	**No. of Citations**	**Partial Sum of Citations (z)**	**Average Citation**	
						Journal	Articles
(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
1	1	20	20	454	454	454.00	22.70
1	2	15	35	245	699	349.50	19.97
1	3	12	47	167	866	288.67	18.43
2	5	11	69	94	960	192.00	13.91
1	6	7	76	143	1103	183.83	14.51
3	9	6	94	235	1338	148.67	14.23
3	12	4	106	87	1425	118.75	13.44
2	14	3	112	111	1536	109.71	13.71
10	24	2	132	166	1702	70.92	12.89
19	43	1	151	61	1763	41.00	11.68

Table 2: A Distribution of Articles based on the Citations Received (Mathematics)

No. of Articles	Partial sum of articles	No. of citations	Partial Sum of Citations	No. of Articles	Partial sum of articles	No. of citations	Partial Sum of Citations
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
1	1	127	127	1	33	21	1177
1	2	70	197	1	34	20	1197
3	5	47	338	2	36	18	1233
2	7	45	428	3	39	17	1284
1	8	43	471	3	42	16	1332
1	9	42	513	3	45	15	1377
1	10	39	552	1	46	14	1391
1	11	37	589	5	51	12	1451
1	12	36	625	3	54	11	1484
1	13	33	658	6	60	9	1538
2	15	32	722	4	64	8	1570
1	16	30	752	5	69	7	1605
3	19	29	839	8	77	6	1653
2	21	28	895	4	81	5	1673
2	23	27	949	10	91	4	1713
1	24	26	975	5	96	3	1728
1	25	24	999	9	105	2	1746
3	28	23	1068	17	122	1	1763
4	32	22	1156	29	151	0	1763

Data Analysis

To begin with, semi-logarithmic curves are drawn for (x Vs. y) and (x Vs. z) for the data given in Table-1 to study the rank distributions of journals and references. They are shown in Figure-1. Also, a semi-logarithmic curve is drawn for the data given in Table-2 (Col.2 Vs Col.4) to study the distribution of articles (references) and citations received. It is given in Figure-2. Figure-1 clearly indicates that the semi-logarithmic curve for x Vs y is close to a straight line and it may be believed that Bradford's law is likely to hold well. The semi-logarithmic curve for x Vs z is close to a typical S-shaped curve indicating that Leimkuhler's model may fit. However, the semi-logarithmic curve for articles Vs citation is close to an S-shape curve (Figure 2). The Figure 2 is discussed in section 3.1

From Table-1 we may note that:

1. Average citation per journal varies from 454 (for the core journal) to 3.21 for the least productive journals;
2. Average citation per journal (for all the 43 journals together) is 41.00.
3. Average citation per article again varies from 22.70 (for the articles in core journal) to 3.21 (for the articles in least productive journals).
4. Average citation per article (for all the 151 articles) is 11.68.

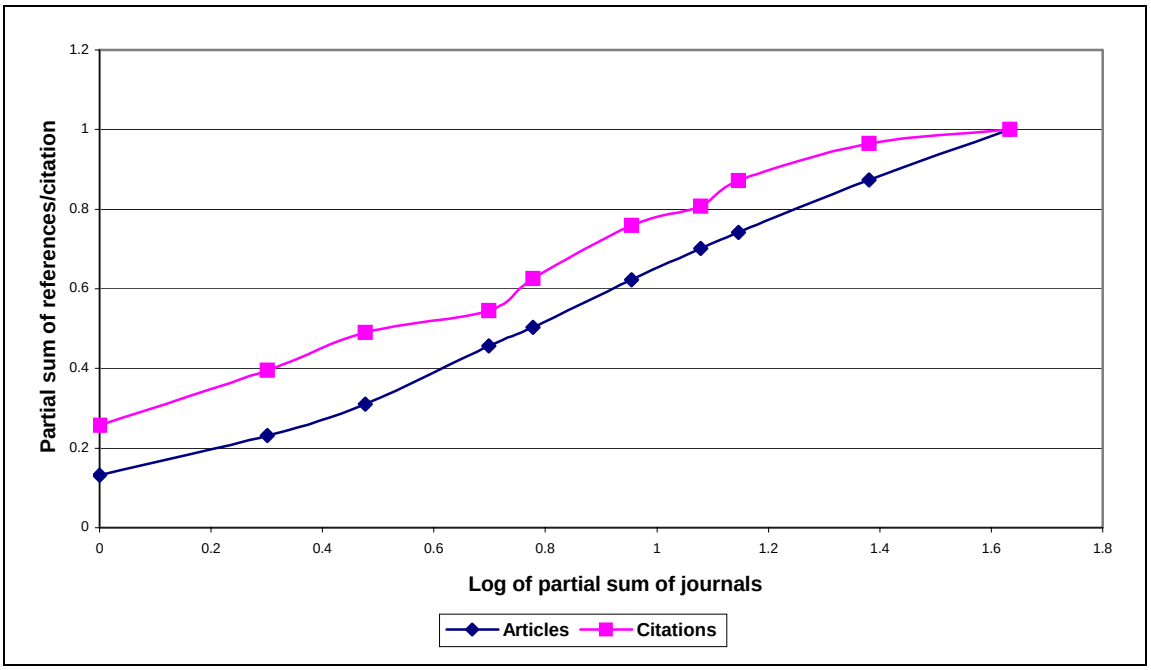

Figure 1: Semi-logarithm curve for the rank of journals Vs Partial sum of references and citations received (Mathematics)

In order to study the relation between y and z (for the data in Table-1), an attempt has been made to fit a log model (as discussed by Basu) and it fits fairly well. Results are shown in Table 3.

Table-3. Observed and theoretical values of citations with relevant statistics

Model z = ay-bylog (y) Method: Least Squares					
	Co-efficient		Std. Error	t-statistic	Prob.
A B	35.456 4.697744		2.791338 0.591693	12.70215 7.939493	0.0000 0.0000
R-squared Adjusted R^2 S.E of regression	0.98499 0.983124 57.04164		Mean dependent Var. S.D dependent Var. F-statistic Prob. (F-Statistic)		1184.6 439.0961 525.3075 0.0000
Observed Value (z)	Estimated (z)	Residual	Observed Value (z)	Estimated (z)	Residual
454	427.656	26.3436	1338	1326.60	11.3998
699	656.386	42.6139	1425	1436.13	-11.1263
866	816.343	49.6572	1536	1488.45	47.5531
960	1074.00	-114.003	1702	1652.36	49.6437
1103	1148.46	-45.4609	1763	1794.80	-31.8023

Similar analyses have also been carried out for another bibliography on Cryptography [2]. Data are given in Tables 4 and 5. The value of R^2 is only 0.85 for the data on Cryptography. This poor value of R^2 is perhaps due to

- A small sample size – only 23 journals covering 56 references are included in the bibliography
- Almost all are most productive references.

Also in this case, we have observed that a most productive article (highly cited article – 824 times, appeared in the journal ranked 3 (-- Physical Review.) Albert Einstein authored this article. However, even for these data, the statistical tests indicate that the model ((y = ay - by log (y)) can be accepted to predict the citations received by a journal.

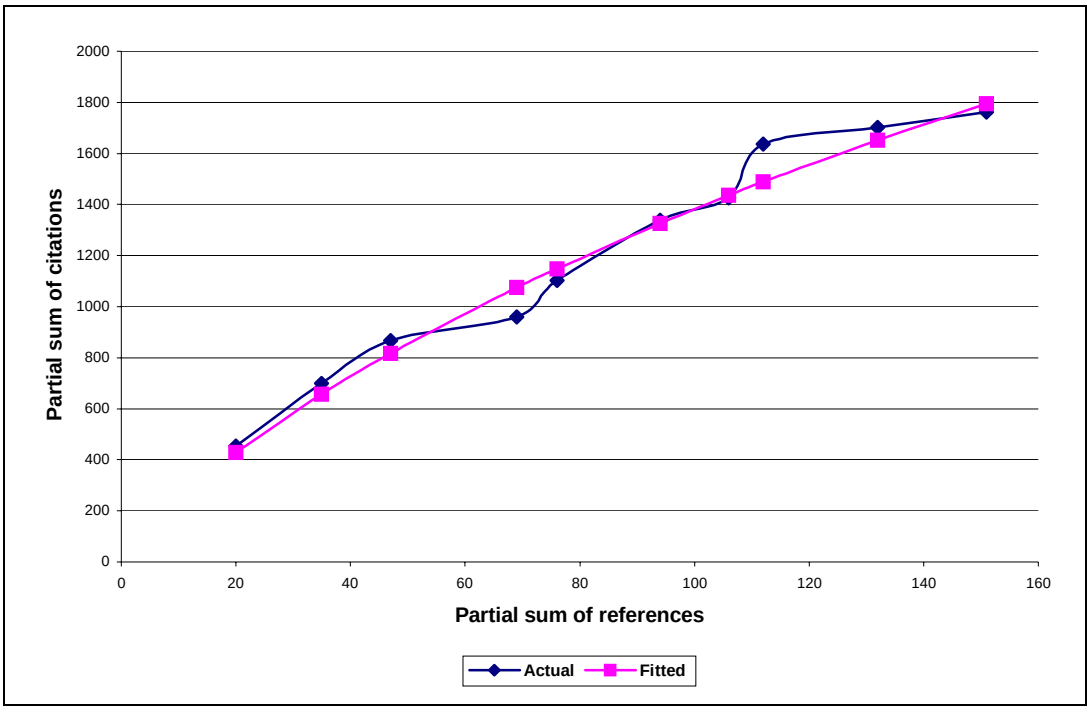

Figure 2: Cumulative curve for the actual and fitted data (based on Basu's model for the data given in Cols (d) & (f) in Table. 1)

Table 4: A Distribution of Journals and References (Cryptography)

No of Journals	Partial sum of Journals	No. of References	Partial sum of References	Partial sum of Citations	Average Citations Per Journal	Per Article
(a)	(b)	(c)	(d)	(e)	(f)	(g)
1	1	13	13	101	101	7.77
1	2	5	18	1481	740.8	82.28
2	4	4	26	1578	394.5	60.69
1	5	3	31	1766	353.2	56.97
7	12	2	45	2188	182.3	48.62
11	23	1	56	3086	134.2	55.11

Table 5: A Distribution of Articles and the Citations Received (Cryptography)

Partial sum of Articles	No. of Citations	Partial sum of Citation	Partial sum of Articles	No. of Citations	Partial sum of Citation	Partial sum of Articles	No. of Citations	Partial sum of Citation
(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1	824	824	11	48	2592	27	60	2996
2	627	1451	12	47	2639	28	11	3007
3	357	1808	13	43	2682	31	27	3034
4	173	1981	14	39	2721	33	14	3048
5	158	2139	15	36	2757	35	12	3060
6	113	2252	17	70	2827	37	8	3068
7	98	2350	18	27	2854	40	9	3077
8	86	2436	19	23	2877	43	6	3083
9	58	2494	21	42	2919	46	3	3086
10	50	2544	22	17	1936	56	10	3086

In cryptography:

i) Average citation per journal varies from 81.64 citations (for the least productive journals) to 101 (for the core journals).

ii) Average citation per journal (for all the 23 journals) is 134.2.

iii) Average citation per article varies from 7.78 (in core journals) to 81.64 citations in least productive journal.

iv) Average citation per article is 55.11.

Thus based on this study, we may conclude that the number of citations (z) received by a journal (on a given subject) can be estimated / predicted the partial sum of articles (y) covered in x most periodicals. Further the author believes that in addition to Egghe's (2003) article on "Type/Token-Taken Informetrics" this study gives sufficient clues to develop 3-dimensional Informetrics. Such a study may be useful to estimate the utility of a journal. Further this study indicate that higher the rank of journals, higher the citations per article as well as higher the citations per journal – in other words the articles in core journals receives more number of citations (22.70 citations per article) than the articles in least productive journals (3.21 citations per article).

A Rank Distribution of References based on the citations received

Bradford in his classical study analyzed the data by arranging the journals in decreasing productivity. In this section, an attempt has been made to analyze the data by arranging the articles (in a bibliography) in decreasing productivity – the productivity is measured in terms of the citations received. The graphical analysis of the data as given in Table 2 (cols (2) & (4) indicates that Bradford's Law holds good even for the rank distribution of articles. If we group the articles into three groups such that each contributes a same number of citations (589 citations) then we have 11, 22 and 77 articles in the first, second and third group respectively. The Bradford multiplier (n) lies between 2 and 2.6. For the data in Table-2, an attempt was also made to fit Lemkulher's model and also Egghe's generalized model. The results are unsatisfactory. Then, an attempt was also made to fit Basu's model

$$F(x) = x - x \log(x)$$

with two free parameters a and b. In Basu's model, F (x) is the number of references that appear in x most journals. In this study, F (x) refer to the number of citations received by x most articles. The results are shown below in Table 6 and Figure 3.

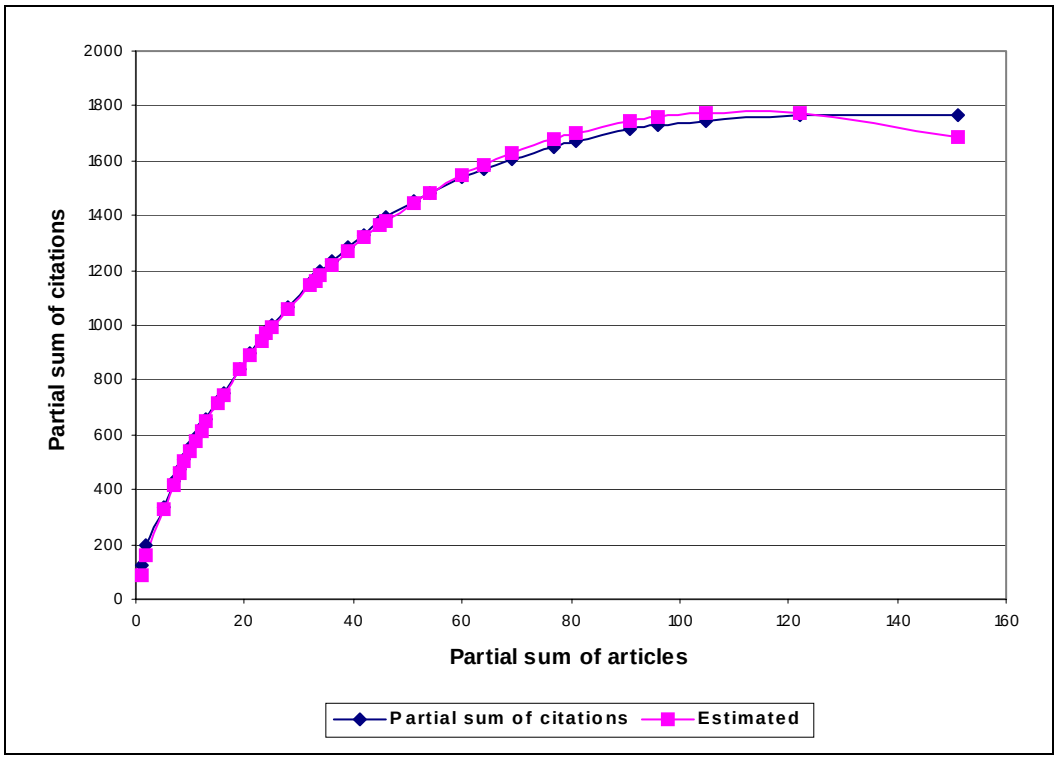

Figure 3: Cumulative curve (with observed and estimated values (y=ax-bx log(x)) Partial sum of articles vs partial sum of citation received.

The results as given in Table 6 indicate that the random hierarchical model fits extremely well to the observed data. The important feature of the model is that it even explains the Groos droop. Based on the figures and data, it is thus conjectured that when the articles are arranged in decreasing productivity, the three or more groups containing same number of citations will have articles in the ration of 1 : n: n^2. The value of n lies between 2 and 2.6.

Table-6. Relevant statistics of fitting of the model F (x) = a x – b x log (x)

Model F (x) = a x – b x log (x) Method: Least Squares				
	Co-efficient	Std. Error	t-statistic	Prob.
A B	0.954725 1.352526	0.008212 0.010686	116.2530 126.5741	0.0000 0.0000
R-squared Adjusted R^2 S.E of regression	0.998108 0.998055 0.012263	Mean dependent Var. S.D dependent Var. F-statistic Prob. (F-Statistic)		0.616404 0.278084 18991.25 0.0000

Thus the authors observe that even ranking of articles yields a good information and this information may be used to select articles for digital libraries – core articles may be subscribed or included and least productive articles may be accessed through online systems. Further, it has been observed that all the articles in the "core zone " are from those journals, which are included in "core zone" in the rank distribution of journals. It makes therefore no difference whether we group the journals or articles. In both the approaches, the Bradford's law is applicable, to identify core journals and thus articles (or core articles and then the journals).

A Size Distribution of Citations

An attempt has been further made here to study the size distribution of citations in a bibliography. The results are shown in Table-7. The average number of citations per article/reference in a bibliography is 11.68. The Kolmogorov-Smirnov test D_{max} (in Table-7) indicates that the distribution of citations follow a negative binomial distribution; the tail is so long that it is unlikely to follow a Lotka's law. It thus indicates that the distribution of citation is a manifestation of success breeds success phenomenon.

Table: 7 A Distribution of Citations received by the articles (Mathematics)

No. of Times Cited	**No. of References**	**Expected Frequencies**	**No. of Times Cited**	**No. of References**	**Expected Frequencies**
(1)	(2)	(3)	(4)	(5)	(6)
0	29	28	22	4	2
1	17	14	23	3	2
2	9	11	24	1	2
3	5	9	26	1	2
4	10	7	27	2	1
5	4	6	28	2	1
6	8	6	29	3	1
7	5	5	30	1	1
8	4	4	32	2	1
9	6	4	33	1	1
11	3	4	36	1	1
12	5	3	37	1	1
14	1	3	39	1	1
15	3	3	42	1	1
16	3	3	43	1	1

No. of Times Cited	No. of References	Expected Frequencies	No. of Times Cited	No. of References	Expected Frequencies
(1)	(2)	(3)	(4)	(5)	(6)
17	3	2	45	2	1
18	2	2	47	3	1
20	1	2	70	1	1
21	1	2	127	1	1
Average is 11.6755		Variance is 264.8550		Standard deviation is 16.2744	
the value of p is 0.0441		the value of q is 0.9559		the value of k is 0.5384	
$D_\alpha = 0.1318$		$D_{max} = 0.069$			

Rank distributions of journals

A journal may be ranked (X) based on the number of relevant articles it contains. It may also be ranked (Y) based on the number of citations received by all the relevant articles it contains. An attempt has been made here to study whether or not there is any difference between X and Y. Two ranked lists (X and Y) are given in Table 8. The Spearman rank correlation coefficient was computed and it is 0.0972515. It indicates that there exists hardly any correlation between X and Y -- two ranked lists are quite different. However it may be observed in Table 8 that the top 10 journals are same in both the list except for one journal!

Table 8: Rank of journals based the number of articles and the citations received

X	Y	X	Y	X	Y
1	1	10	16	10	31
2	2	8	17	11	32
3	3	9	18	11	33
6	4	11	19	11	34
7	5	11	20	11	35
4	6	10	21	10	36
9	7	11	22	10	37
7	8	11	23	11	38
10	9	10	24	11	39
8	10	10	25	11	40
7	11	11	26	11	41
10	12	11	27	10	42
8	13	11	28	11	43
5	14	11	29		
10	15	11	30		

Conclusion

Two bibliographies -- a bibliography for automarphic and modular forms, L functions and representation theory and a bibliography of quantum cryptography – were used to collect the data on rank distribution of journals, articles and distribution of citations. This study indicates that

- Number of citations received by a journal (z) can be predicted using the number of articles (y) it contains using a model z = ay - by. log (y)
- The rank distribution of articles closely confirm to Bradford's law; in other words, based on this study, we can argue that when the articles are arranged in decreasing productivity, the three or more groups containing same number of citations (utility) will have articles in the ratio of 1: n: n^2 …
- Further rank distribution fits Basu's Model z = ax – bx log (x) very well.
- The size distribution of citations in a bibliography (x: the number of citations; f (x): the number of references with x citations) follows a negative binomial distribution.

- The average citations per article is very high for those articles published in highly productive journals (22.7) as compared to the to those articles which are published in least productive journals; this also implies that the average citations per core journal is very high as compared to the least productive journal – **higher the rank of a journal, higher the citations it receives, particularly in mathematics.**
- The two ranked lists of journals (based on articles and citations) are quite different; however, among the most productive journals there is no much difference**.**

Bibliographical References

A Bibliography for Automorphic and Modular forms, L-functions, and Representation Theory. http://www.math.umn.edu/~garrett/m/b/bib.html

A Bibliography of Quantum Cryptography. http://www.cs.mcgill.ca/~crepean/CRYPTO /Biblio-QC.html

Basu, Aparna. (1992). Hierarchical Distributions and Bradford's Law. *Journal of the American Society for Information Science*. 43, 7; p.494-500.

Bradford, S.C. (1934). Sources of Information on Specific Subjects. (*Engineering*, 137, 85-6)

Brookes, B.C. (1969). Bradford's Law and the Bibliography of Science. (*Nature*, 224, 6,953-6.)

Egghe, L. (1990). New Bradfordian Laws equivalent with old Lotka laws, evolving from a source item duality argument. In L. Egghe and R Rousseau, Informetrics 89/90, *Proceedings of the 2nd International Conference on Bibliometrics, Scientometrics and Informetrics* (p. 79-96), Amsterdam, Elsevier.

Egghe, L. (2003). Type/Token – Taken Informetrics. *Journal of the American Society for Information Science and Technology*, 54, 7, p.603-610.

Leimkulher, F.F. (1967). The Bradford distribution. *Journal of Documentation*, 23, 3, p.197-207.

Vickery B.C. (1948). Bradford's Law of Scattering. *Journal of Documentation*, 4,3, 198-203.