WHO REFERS TO WHOM:
A Study of Research References and the Relationship between Research Reports and Final Publication

By

Samuel McCarthy, Martin Shubik, and Jianfeng Yu

January 2003

COWLES FOUNDATION DISCUSSION PAPER NO. 1396

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY

Box 208281

New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Who Refers To Whom: A Study of Research References and the Relationship between Research Reports and Final Publication

Samuel McCarthy, Martin Shubik, Jianfeng Yu*

January 2003

Abstract

The size and style of referencing for a large sample of 60 years of publications of the Cowles Foundation are examined. The influence of computerization is considered. Self-referencing is noted and some observations are made on the costs and distribution of research papers.

1. Introduction

This study involves the data based on 60 years of publications (1943–2001) of Cowles Foundation Papers (CFP) and 46 years (1955–2001) Cowles Foundation Discussion Papers. The first investigation concerns the distribution of the dates of reference, to see whether there is a systematic change in these distributions over time. The characterization of the nature of the tails of the distribution is considered.

The data were sampled from 1350 Cowles Foundation Discussion Papers. Ten papers were sampled randomly from the discussion papers for each year. For each paper sampled, the date when it was finished and the dates of the references in the paper were noted. One of our concerns is with characterization of the distribution of the time difference between the discussion papers and their references. The sampling of ten papers in each year is somewhat skimpy but we were limited by time and costs. For the 1955 and 1956 data, the total number of discussion papers was less than 10. Hence for those two years, we sampled all the discussion papers.

* The authors wish to thank Glena Ames and Art Trager for assistance.

An example of a sample datum is as follows: (one discussion paper)

Date of Discussion Paper	Date of All References
1955	1953
1955	1950
1955	1954
1955	1955
1955	1952

After recording the dates of the discussion paper and its references, we calculate the difference between the two columns to get the time difference between the discussion paper and its references. We focus our interest on the distribution of this time difference.

Secondly, we investigate the relationship between the publication of Cowles Foundation Discussion Papers (CFDP) and Cowles Foundation Papers (CFP). Has it remained constant over time? What is the nature of the time lag between publications of a CFDP and the resultant CFP? Sample data look as follows:

Date of CFDP	Date of Publication of CFP
1955	1958
1956	1957
1956	1962
1956	1957
1956	1957
1956	1958
1956	1957

2. Simple Results

First, we note the aggregate distribution of the time difference between the discussion papers and their references. The histogram below is based on all sampled papers. We also find that the one year old reference is the most frequent, the two year old reference is the next and the same year reference is the third most frequent. This appears to be plausible. Individuals often obtain ideas from recent papers and refer to those papers.

It usually takes a long time to publish a paper even after it is accepted by a journal. Thus, individuals and people may refer to a to-be-published paper having seen an extant but not necessarily dated version. In our sample as there was only one such paper, we omitted it from the estimate of the distribution for the time difference.

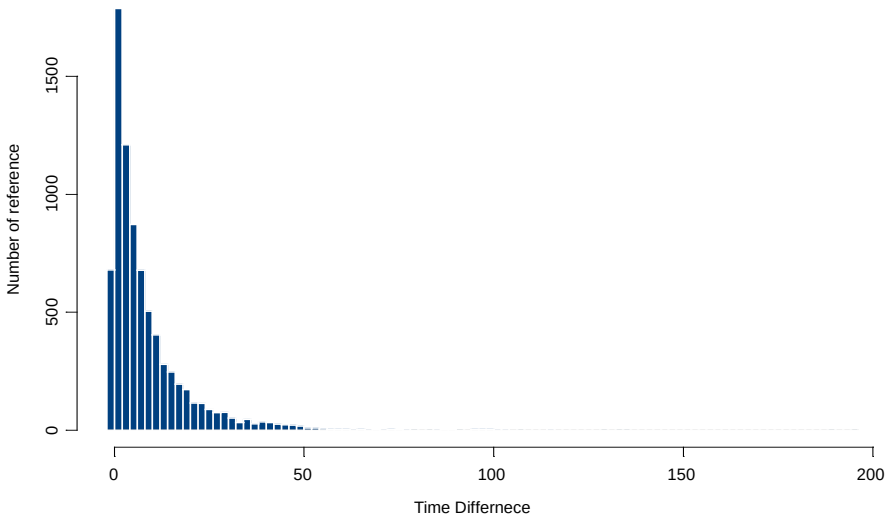


Figure 1. The Histogram for the Whole Dataset

Some references are about 200 years old, but almost all the references are less than 50 years old. We denote the time difference random variable as T. We first test to see if the distribution of T is similar to a Poisson distribution or if $T + r$ is similar to the negative binomial distribution (which can be considered as a modification of a Poisson, where r is some integer constant). We find that the sample mean is 9.1 and variance is 151 (if we only consider those samples whose time differences are less 100, the mean would be 8.2 and variance would be 84). However, the mean and variance of the Poisson distribution should be the same. Hence, this rules out the Poisson distribution.

We try a transformation and use a Chi-square test for the hypothesis that $T + r$ belongs to the negative binomial family NB(r, p). The asymptotic result requires us to find the MLE of r and p. We can estimate r and p roughly through the mean and variance expressions (the method of moments). Note that r must be an integer. We get $r = 1$, and $p = 0.1$ based on the samples whose time difference is less than 50. A plot that compares the empirical distribution and this estimated negative binomial is given in Figure 2. It is seen that it appears to fit the empirical distribution well except for the first several points. However, the Chi-Square goodness of fit test shows that the difference is significant.

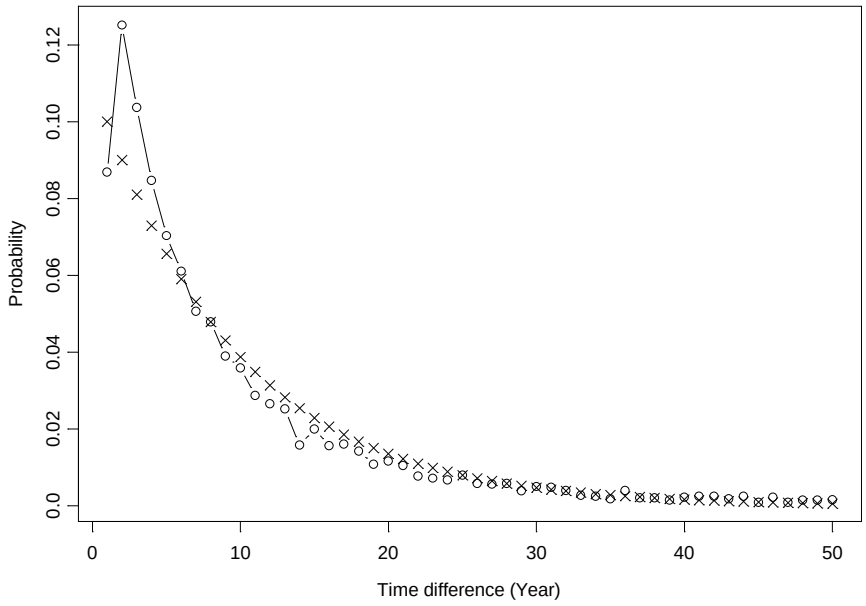


Figure 2. The Empirical and Fitted distribution for the Time Difference

The following is the number of references for the ten papers for each year. It is seen that there is a significant increase in the number of references with a jump around 1984, 1985.

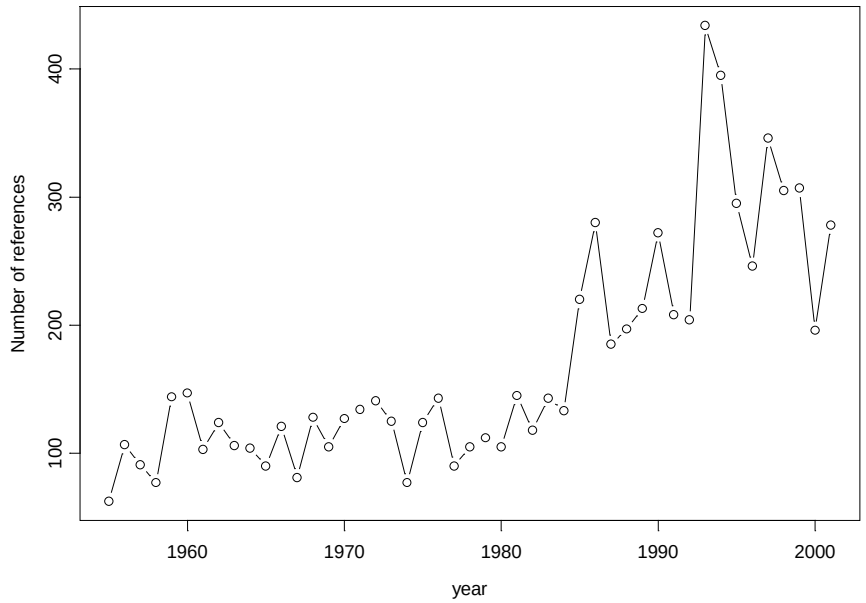


Figure 3. The Number of References in the Ten Sampled Papers over 1955–2001

Before around 1985, people could obtain references only by recording by hand or possibly by photocopying, but since that time, increasingly it has become possible to obtain and review

references online. This change in the referencing nature of scholarship may offer a plausible explanation for the change in the volume of references from 1985 onward.

In each year, we use the time difference between the discussion paper and its references in the sampled ten papers to obtain an empirical distribution of the time difference for that year. We use this empirical distribution in turn as an estimator of the time difference distribution for that particular year. First, we want to see if there is any trend in the means of those distributions. Hence, we regress the mean value on the publication date (Year). We wish to examine if the coefficient of the variable "Year" is zero, negative or positive. We find that the estimation of the coefficient for Year is 0.113, indicating that the time difference between the discussion paper and its references increases with time. Since the estimation of the coefficient is small and the positive value could come from the random effect. We test the hypothesis that the coefficient of Year is zero. To test this hypothesis, we use a *t*-test. If the null hypothesis holds, then the *t*-statistic will be very small. For our problem, the *t*-statistic has a *t*-value of 5.7, the corresponding probability that the absolute value of the *t*-distribution is greater than 5.7 is 9.1e-7. Hence, there is only a very small chance for a higher absolute *t*-value than 5.7, which implies that we can reject our null hypothesis.

Table 1

Coefficient	**Estimated Value**	**Standard Error**	***t*-value**	***p*-value**
Intercept	–215.56	39.37	–5.5	1.9e-6
Year	0.113	0.02	5.7	9.1e-7

Indeed, from the above table, we see that the *p*-value of "Year" is very small and its coefficient is positive, implying that the mean time difference between discussions papers and their references increases significantly.

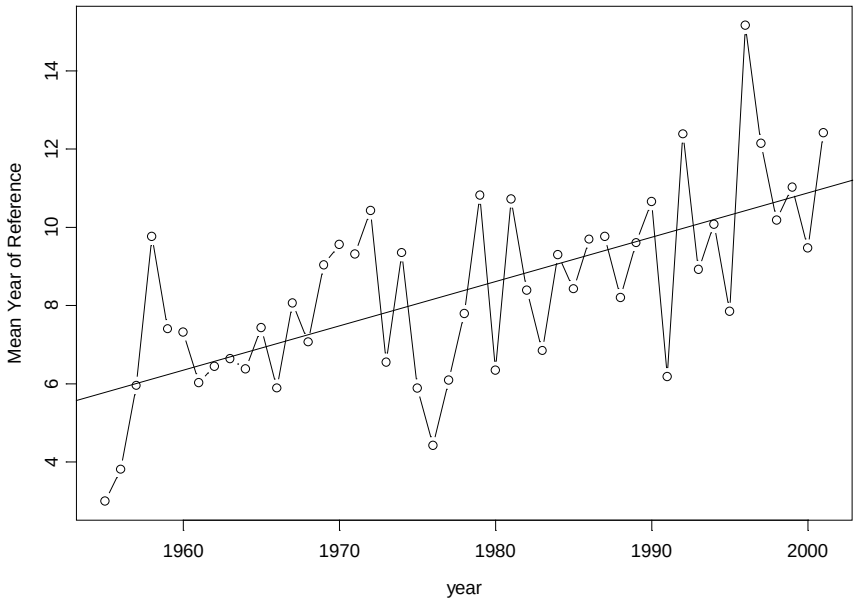


Figure 4. The Mean of Time Difference in Each Year over 1955–2001

Since the mean can be influenced greatly simply by the outliers (some of references are over 100 years old) while the median is much more robust, we plot a median-year graph. It is seen that the median is increasing significantly with time passing.

Table 2

Coef.	Estimated Value	Standard Error	*t*-value	*p*-value
Intercept	–154.35	29.908	–5.2	5.4e-6
Year	0.080	0.015	5.3	3.2e-6

From the above table, we know the coefficient of "Year" is significantly positive since the *p*-value is extremely small. Hence, the time difference between discussion paper and its references is statistically significant increasing with time.

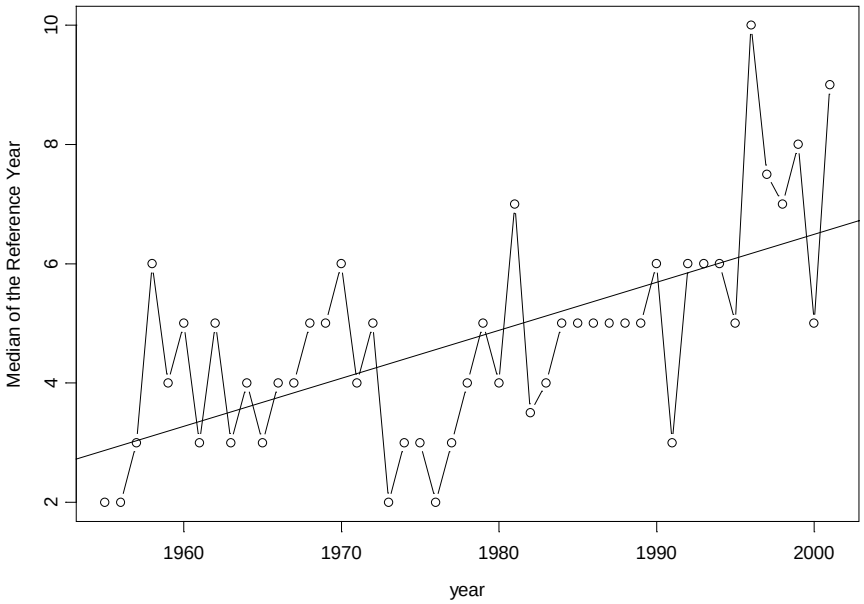


Figure 5. The Median of the Time Difference for Each Year over 1955–2001

Much of our concern in this study is descriptive of the referencing. Causality is difficult to establish, however we have several conjectures and "guesses" which we do not formally establish. But note, from late 1960s to 1980s, several seminal papers were published, which became classical references for later papers.

Furthermore it is likely that as the personnel turn over in an institute such as the Cowles Foundation is relatively slow and there is a tendency to self-reference and cohort group reference; as individuals age their references might also age.

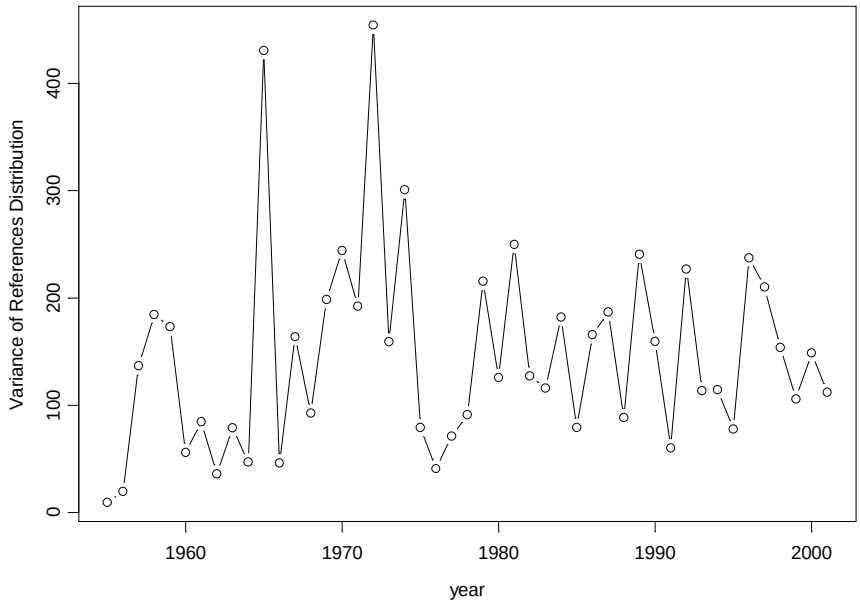


Figure 6. The Variance of the Empirical Distribution of Time Difference for Each Year over 1955–2001

The above plot is a graph of the variance of the empirical distribution for each year. Because there are some outliers (those references older than 100 years), some of the variances are very large. Furthermore the variance becomes very unstable over the years. Since L^1 distance (Total variation) is robust we use the L^1 distance to see if there is any trend in this distance with time.

As we have the estimator of the distribution of time difference for each year, we can calculate the L^1 distance (Total Variation) of any two successive year's distribution. we find that the L^1 distance between the two successive years is decreasing, but it is just slightly significant.

Table 3

Coefficient	Estimated Value	Standard Error	*t*-value	*p*-value
Intercept	4.814	2.145	2.25	0.03
Year	–0.002	0.001	–1.98	0.05

From the above table, we find that the p-value is 0.05, and the coefficient of "Year" is a small negative number. Hence, the L^1 distance is slightly significantly decreasing.

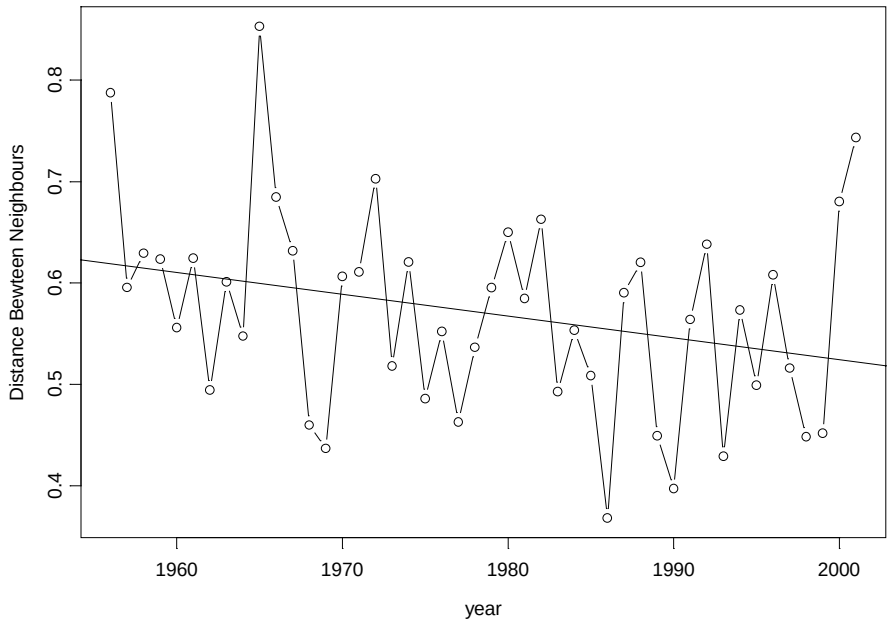


Figure 7. The L^1 Distance between Successive Years over 1955–2001

We use all the sampled papers to find an aggregate empirical distribution for the time difference between the discussion papers and their references; then calculate the relative L^1 distance between any empirical distribution for any single year and the aggregate empirical distribution. Finally, we can see that the relative L^1 distance decreases significantly, so the distribution of the dates of the references converges to the average distribution, which indicates that the distribution of the dates of references is becoming more stable.

Table 4

Coefficient	Estimated Value	Standard Error	*t*-value	*p*-value
Intercept	9.736	2.340	4.161	1.408e-4
Year	–0.005	0.001	–3.958	2.656e-4

The following graph shows the fitted regression line in the L^1 distance. It is seen that the slope is negative.

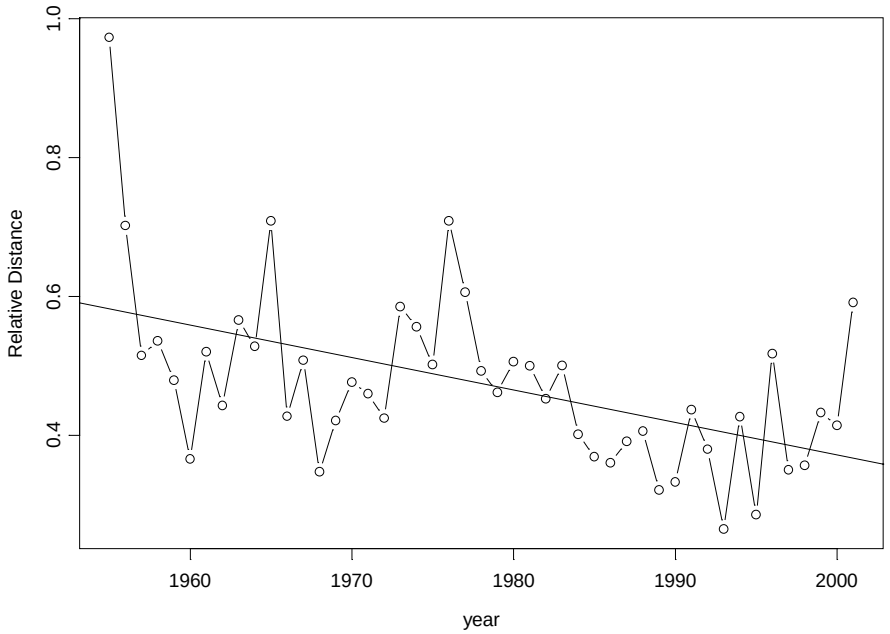


Figure 8. The Relative L^1 Distance for Year 1955–2001

We already know that the variance will be influenced easily by outliers that have large time differences between the date of a discussion paper and its references. However, entropy cannot be influenced by these outliers. Furthermore, entropy is also a good measure of variation of distribution. For a discrete distribution {Pi, i = 1, 2, 3, ..., N}, the entropy of this distribution is the summation of -Pi*log(Pi). We calculate the entropy for each single empirical distribution, and we find that the entropy of these single empirical distributions is significantly increasing. From the former analysis, we see that the distribution of the time difference between discussion papers and their references will "converge" stably to some distribution. Note that the final distribution has heavier tail, which implies that it has larger entropy. Hence, the entropy should increase over time.

Table 5

Coefficient	Estimated Value	Standard Error	*t*-value	*p*-value
Intercept	–33.370	4.741	–7.039	8.920e-9
Year	0.018	0.002	7.644	1.144e-9

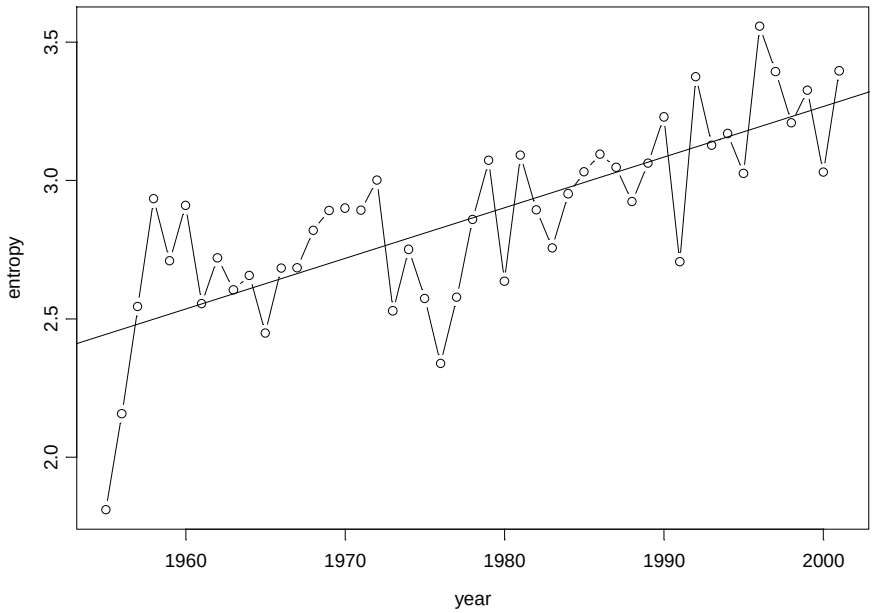


Figure 9. The Entropy of the Empirical Distribution for Each Year over 1955–2001

From the above Table 5, it is seen that the *p*-value is extremely small. Hence the entropy is statistically significantly increasing over the years.

Table 6

Coefficient	Estimated Value	Standard Error	*t*-value	*p*-value
Intercept	–9.595	1.242	–7.725	0.000
Year	0.005	0.001	7.922	0.000

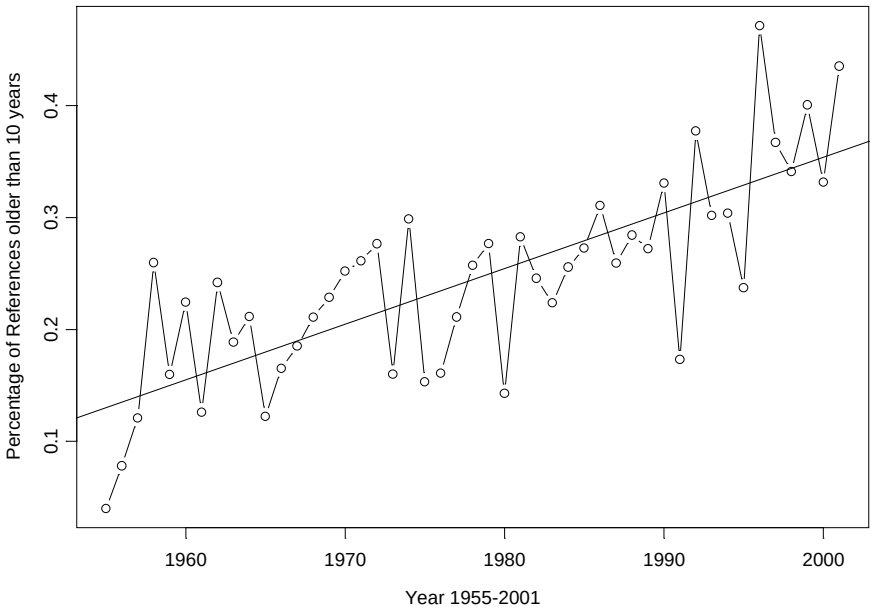


Figure 10. The Percentage of References Older than 10 Years in Each Year over 1947–2001

For the relationship between Cowles Foundation Discussion Papers (CFDP) and the Cowles Foundation Papers (CFP), we sample the data from 1955 to 1994. However, because some of the CFDPs after 1995 will possibly become CFPs later, the data after 1995 may be inaccurate if we use them.[1] There are some discussion papers that are published as a CFP more than ten years later. They can be regarded as outliers and may be deleted in the subsequent analysis, but they also may contain a commentary on the refereeing processes in economic journals.

The following Figure 11 plots the percentage of CFDPs that become CFPs versus time (year). A qq-norm plot (Figure 12) is also given. It shows that there is no particular relationship between these percentages over time. It appears to be white noise. This may be reasonable. Some individuals publish their discussion papers in books; some people do not publish their discussion papers in any journal; and some do not publish their discussion papers as a CFP even though it is published in some journal.

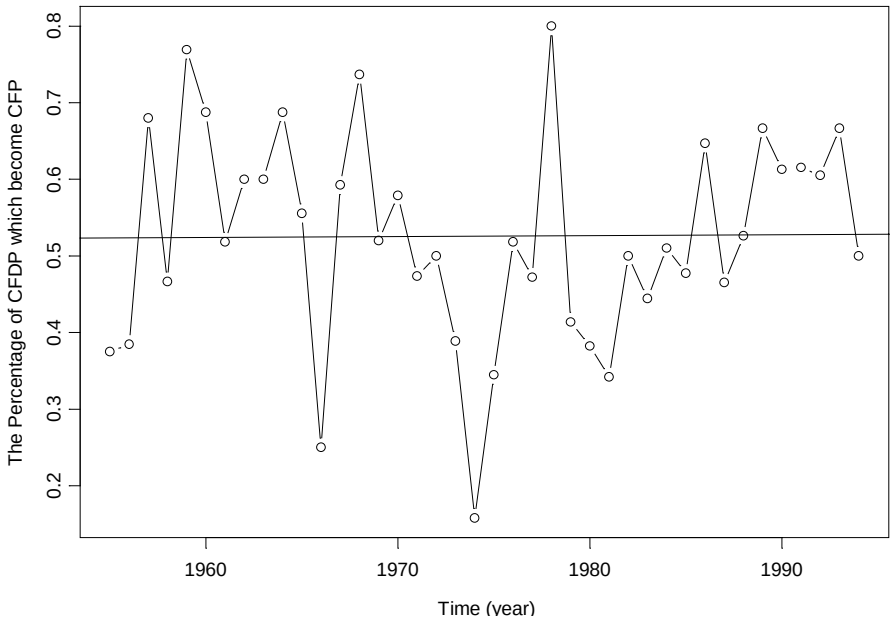


Figure 11. The Percentage of CFDPs Which Become CFPs Later for Each Year over 1955–1994

[1] There are even a few CFDPs from the 1980s still tied up in journal refereeing, but the number is small enough to be safely omitted.

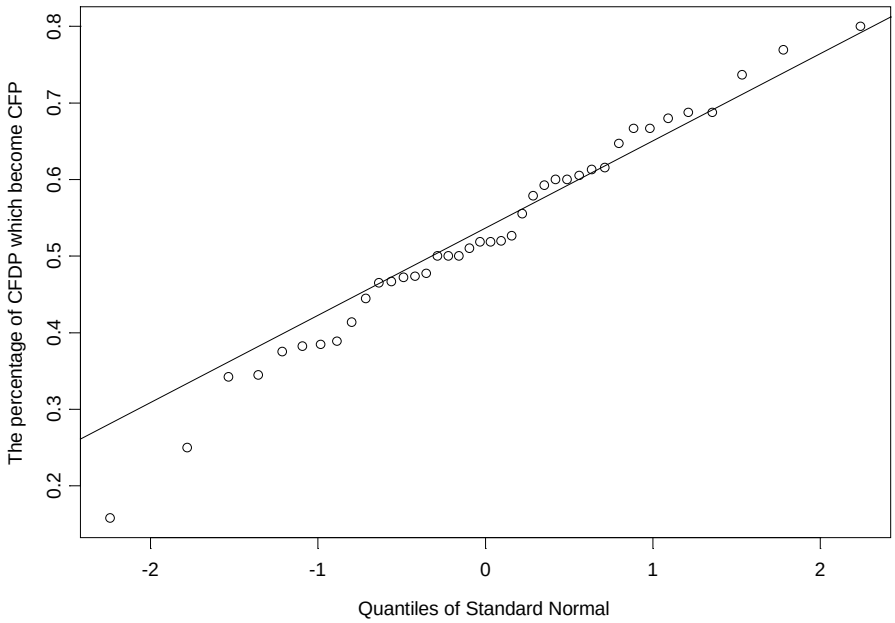


Figure 12. The qq-norm Plot for the Percentages

From the following graph, we can see that the number of publications of CFDPs as CFPs has an increasing trend.

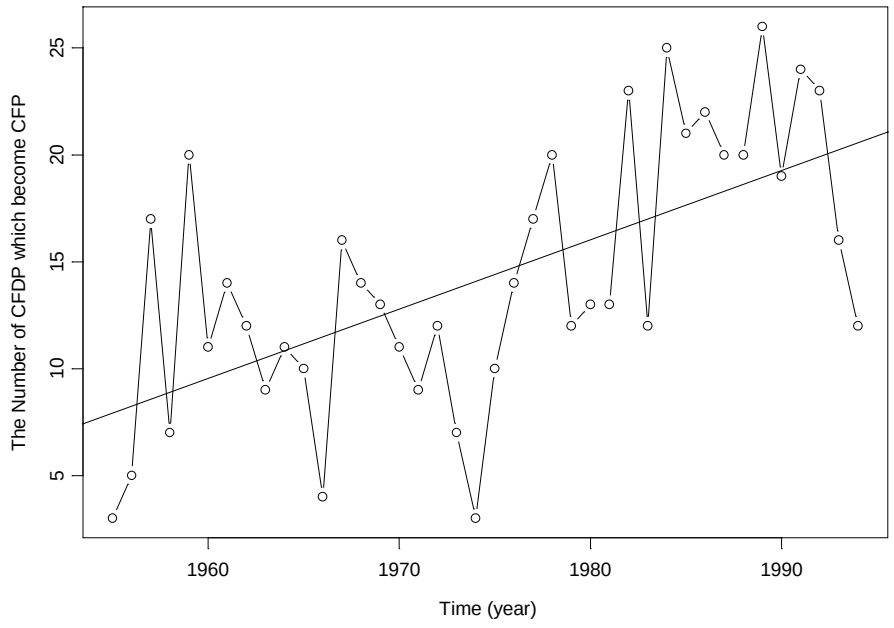


Figure 13. The Number of CFDPs That Become CFPs in Each Year from 1955–1994

We can calculate the average time lag between CFDPs and CFPs for each year. We find that the time lag between the publications of CFDP and the resultant CFP has no significant trend

over time. It is very messy like white noise. If we use a linear regression, we find that the hypothesis of the coefficient of the time being 0 is accepted (*p*-value is 0.69).

In the first year there were few discussion papers that became CFPs and the average time lag was large, we can treat the first observation as an outlier. We can delete it, and do regression again. However, it is still insignificant (the *p*-value is 0.07).

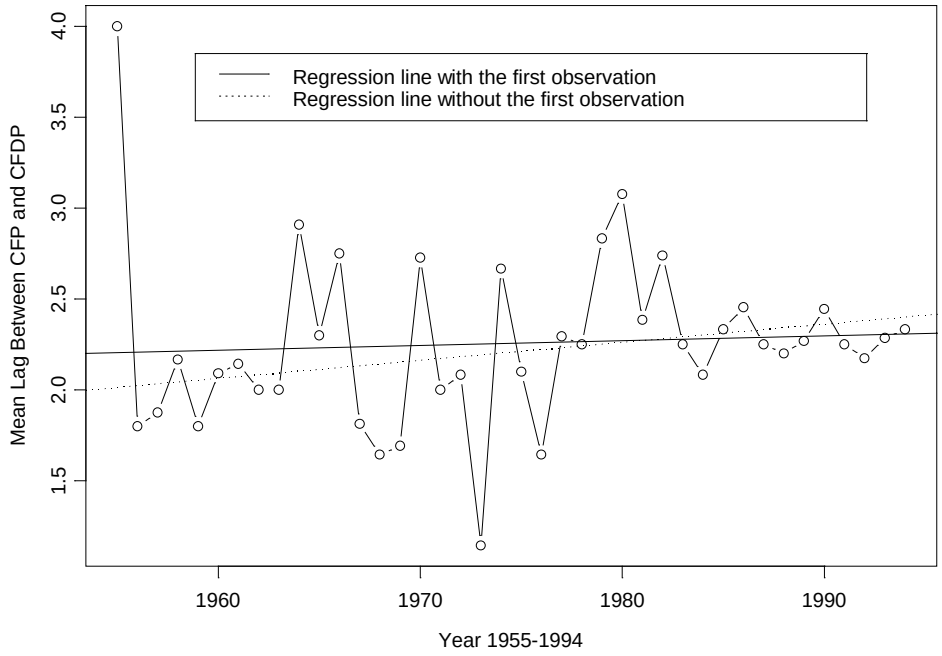


Figure 14. The Average Time Lag between CFP and CFDP for Each Year over 1955–1994

From the above analysis, we can conclude that people usually publish their discussion papers in two to three years, but there are some random effects (such as the refereeing procedure and production delays) which will lead to discussion papers being published many years later. Those random effects are almost independent. The average time lag falls around somewhat more than 2 years, but with a considerable random component.

The following qq-norm does not appear too promising. However, as the number of observations is not large, it is still quite normal. We plotted QQ-norm plots for 40 random number generated from normal distribution twenty times. We found that some of the QQ-norm plots are even worse than Figure 15. Hence, it is reasonable to infer that the time lag is basically white noise. We also note that the time lag becomes much more stable from 1980 on. The average time lags for 1995–1998 are between 1.5 and 2.0. It appears that the average time lag

after 1980 has stabilized at between 2 and 2.5 years. We have no clean hypothesis as to why there is this apparent stability.

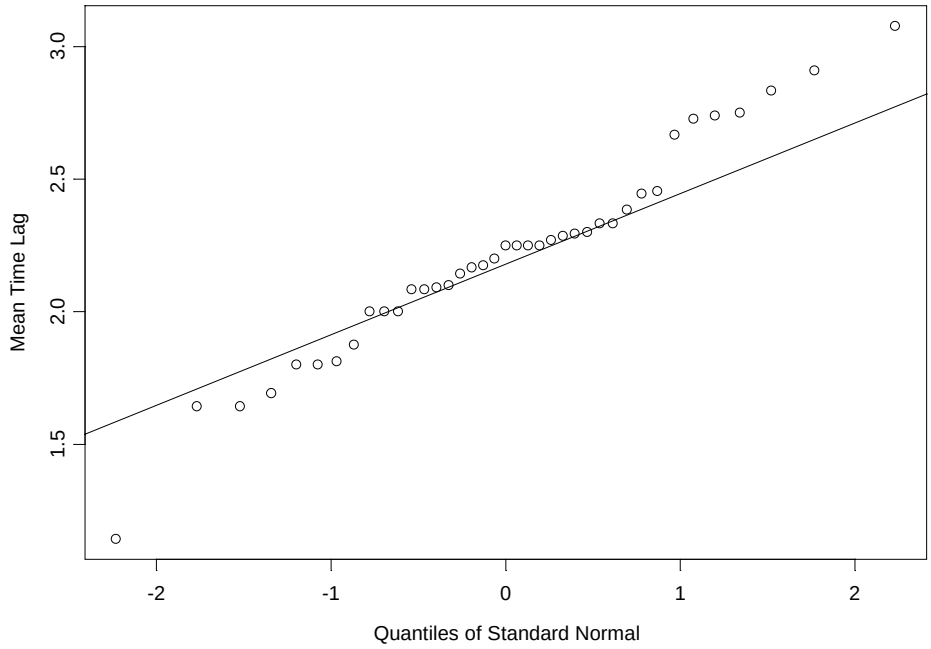


Figure 15. The QQ-norm Plot for the Time Lag

3. Discussion

We have found that the average time difference between the original paper and its references has an increasing trend with time, and the distribution of the reference date is somewhat similar to a negative binomial distribution. Furthermore, the distribution of the reference dates appears to converge to some stable distribution and the "limit" distribution has a somewhat heavy tail, hence has larger mean and variance.

The average number of references per paper in the 1960s to 1985 was around 10 and since then from 1985 to the present is around 25.

As to the relationship between the CFPs and CFDPs, we found that the percentage of CFDPs that become CFPs later has no particular trend with time. Its behavior is like "white noise." The time lag between publication of a CFDP and the resultant CFP has also no significant trend with time. It takes around the same average time to publish a CFDP today as before. But since 1980s, the average of the time lag has become more stable.

We may ask the inverse question "Are there CFPs which never were CFDPs? The answer is yes, many.

Figure 16 shows the number of publications of CFDPs over the years and the number of CFPs.

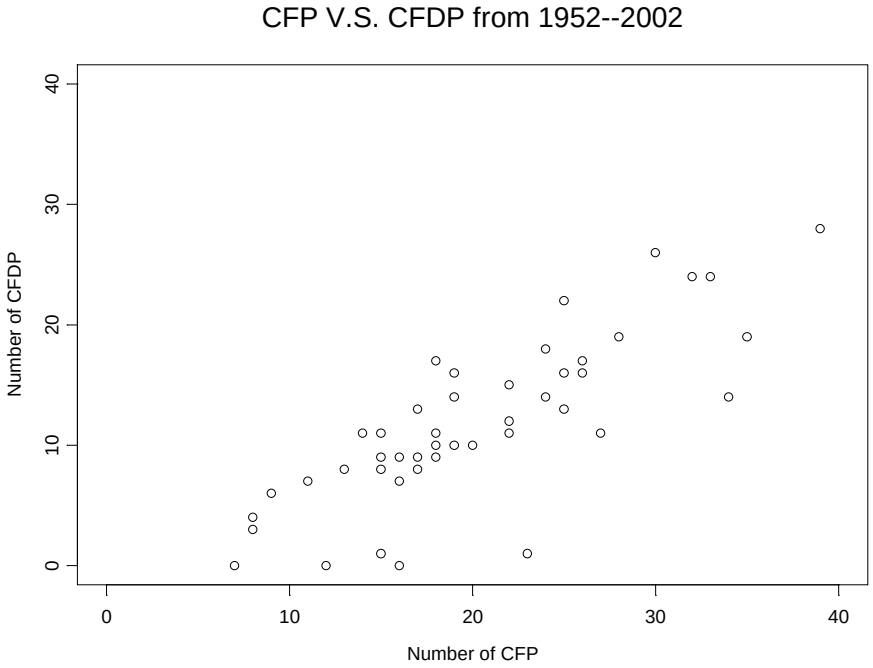


Figure 16: Number of CFP vs. Number of CFDP from 1952–2002

Of the 1040 CFPs to date, CFPs 1-98 were produced when Cowles was in Chicago. The remainder, 119 – 1040 = 946. Of this number 617 were previously CFDPs, or approximately 70.8% but this number could be somewhat adjusted when one considers that 20 CFPs had two CFDPs each as antecedents.

4. The Tail of the Distribution and Self Reference

Although our prime concern was to characterize the overall size and distribution of references we had some further conjectures, which we note here. They are not particularly deep, but to some extent it often worth trying to check the obvious, to see if in fact it is obvious. In particular we expected that references to articles more than 40 or 50 years old would be references to the work of those individuals who had been sorted out by time to be the great individuals in the field. We also felt that it was worth looking at the nature of self-reference. Our

expectation was that it would be different in different fields of economics, with it being the least in applied topics such as macroeconomic policy and the highest in controversial theory.

4.1. The Tail of the Distribution

From the first paragraph of the result in this article, we already know that the tail distribution is close to the tail of Negative Binomial distribution with the estimated parameter.

More specifically, we examined the references to articles or books published over 50 years earlier than the CFDP quoting them. There are in total 97 of these references in articles in our sampled papers over the 47 years. (There are 12, 26, 35, 47, 62, and 97 articles that are 100+, 90+, 80+, 70+, 60+ and 50+ years older than the articles we sampled respectively.) We examined the references in about 200 CFDPs. The authors of the 24 articles that were cited over 50 years later were:

> Jevons, W.S., Bertrand, J. (twice), Edgeworth, F.Y. (4 times), Walras L., Ramsey, F.P. (twice), Fisher, I. (6 times), Marshall, A., De Tocqueville, A., Hamilton, A., Tinbergen, J., Mahan, A.T., Smith, A., Zeuthen, F., Kermack, W.O., and A.G. McKendrick.

It is easy to see that for almost all of these articles, the authors are authorities in their fields. Hence, it lends plausibility to the fairly obvious conjecture that within a century much of the work has been absorbed leaving only a few major authorities.

4.2. Self-references

As our sample size consists of only eight individuals we make no pretense at statistical significance. We selected eight members of the Cowles Foundation in different fields. In particular one in macroeconomics, two in econometrics, one in programming and economic theory, one in mathematical economics, one primarily in economic theory and game theory, one in finance and one in general microeconomic theory

By looking at the self-references in Cowles Foundation papers, we explored the possibility of different referencing and the nature of the subject matter. For instance, we would expect an abundant amount of cohort references available to those writing in macroeconomics. However, we would not expect to have the same abundance available to an economist writing in mathematical microeconomic theory or econometrics.

We would then expect the frequency of self-references to increase as the subject matter becomes more abstruse and technical. On this basis we expected Macroeconomics as the lowest; with mathematical economics; econometrics; programming and economic theory; and game and economic theory as considerably higher.

We sampled a certain amount of papers in accordance with the economist's output over the duration of publishing. We then counted the total amount of references and divided the two. We found the following data:

Economist	Sample Size	Number of Self References	Total References	Self/Total
Macroeconomics	9	10	218	4.59%
Theory (Math econ)	15	42	183	22.95%
Programming, theory	13	23	103	22.33%
Econometrics	12	56	263	21.29%
Theory (Math econ)	12	28	255	10.98%
Micro theory	7	4	255	1.56%
Econometrics	14	47	441	10.66%
Finance	13	43	419	10.26%

The study of self-referencing involves many dimensions which we do not cover here. For example there may be inbred cliques and individual personality may influence self-referencing. Our comments here may be viewed as a first cut at observing the size of self-referencing and considering differences in economic sub-disciplines.

5. Costs and the Distribution of Research

We have already observed that technology has already had an apparently large influence on the nature of referencing. Of far more importance to research has been the cost and speed of the dissemination of new results. In particular when the first CFPs were produced in 1942 a handful of economic journals existed, international travel was slow, conferences were far less frequent than now and "snail-mail" (though possibly better than now) meant that the dissemination of new results, questions or ideas took weeks if not months.

Currently we live in a world of e-mail and the web. At Yale there is a joint subscription to the Social Science Research Net. The fee covers the Cowles Foundation, the Economics department and the Economic Growth Center. The subscription fee is $13,500 which is an overhead, while the marginal cost of publishing a downloadable paper is the cost of having a

computerized version of the paper posted on the web. Currently virtually all researchers prepare their own papers in some partially or fully acceptable computerized form. Little extra computer and secretarial professional work is required. The overhead is in part the time of one skilled mathematical document producer.

The Cowles Foundation still produces physical CFDPs and the costs are estimated at 6 cents per page. A sample of 11 papers was taken from the 2001 CFDPs to calculate a crude average of cost based on the average size and run of the papers. The average length was 40 pages and the run until recently was 100 copies. This gives $240 per CFDP not counting postage.

The reprint cost for a CFP vary in price from approximately $300 to $600 per 100. The common order for a CFP is 100.

In 2001 63 papers were published as CFDPs, thus the estimated cost of publishing CFDPs is $15,120. In that year CFP reprint costs were approximately $4,000. Using these rough estimates, a case can be made for more reliance on electronic publishing and less on the physical production of CFPs and CFDPs. In particular, not only does the downloading and printing of papers by those who wish to read them assign the postage and paper costs to the individual reader, these relatively small nuisance costs together with the downloading information provide a better indication of the readership than is currently available. Furthermore, rather than carry a CFDP inventory, the CFDPs are produced to order.

6. Concluding Remarks

The computerization of both article search and referencing together has changed the task times in scholarship. The speed in which preliminary papers can be put up on the net has changed diffusion rates for research ideas. Furthermore finally published papers, even though they appear in hard copies of journals, are more and more being posted in a downloadable form. These developments are changing not merely the economics of publication, but are changing the possibilities for research on large data banks of information.

As many of the CFDPs and CFPs considered here were not yet in computerized form the research for this article represents a blend of old style highly time consuming data assembling prior to obtaining a large enough sample size. When computerization is complete, we will be in a position to utilize 100% or near 100% samples and some useful insights concerning trends in scholarship can be obtained.

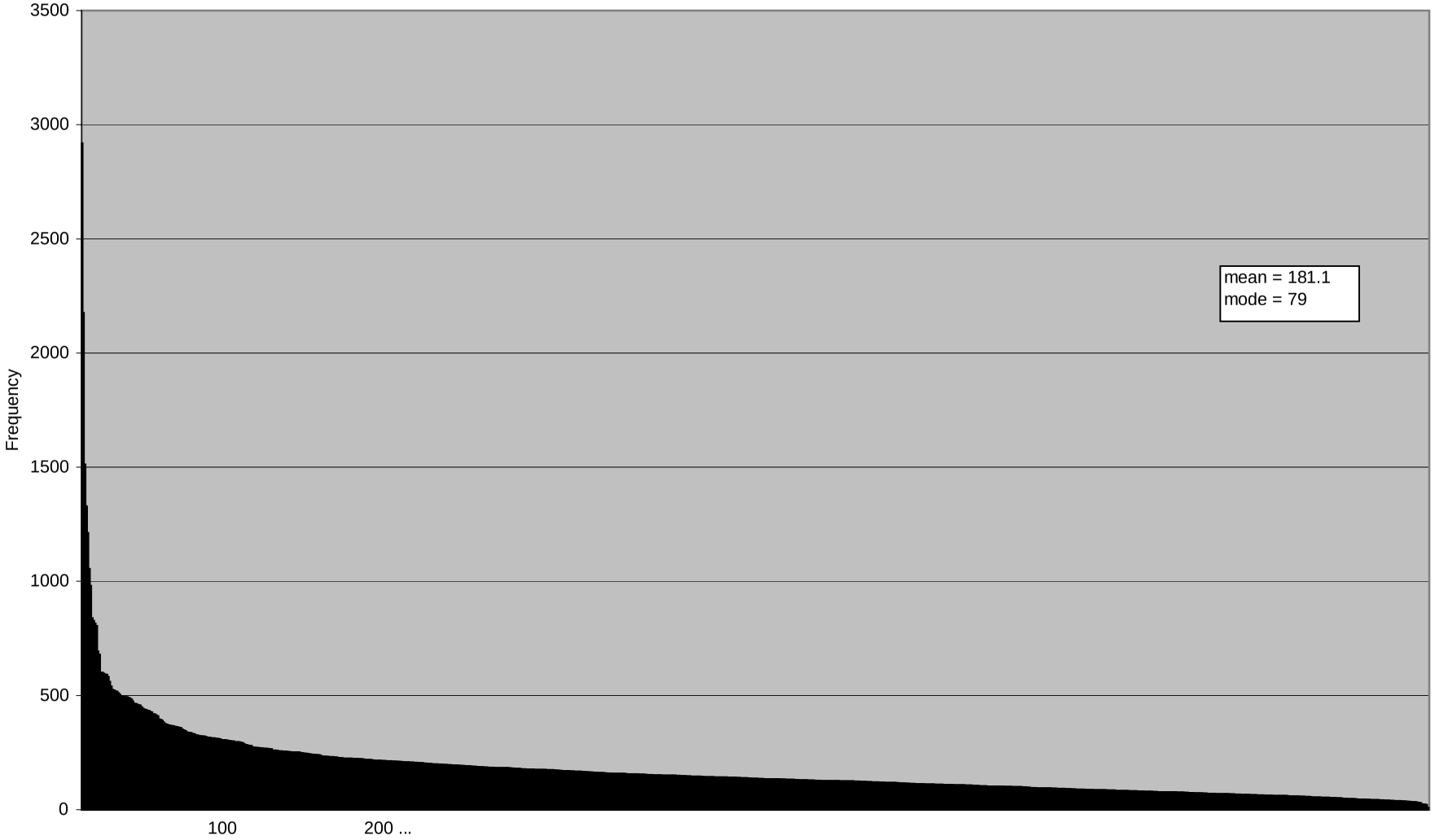
The ability to assemble larger and larger samples of numerical date brings with it new hazards as well as new gains. In particular, although we are now able to count website hits and receive popularity poll information as to whether your recent web-published paper is in the top ten, the meaning and use of such information must be approached with circumspection. Before we become too infatuated with what the numbers are and how we can massage and manipulate them we must be concerned with what the numbers mean.

Indices are a social as well as a scientific construct; as such we must be concerned about the biases of those who construct them and the purpose to which they are put. Popular and easily understood articles may easily attract the most web hits. They may well be ephemera, but can be misused by a business school dean or department chairman as proof to the public that their scholars are known. It may take us another 10 to 30 years before we can check with some confidence concerning the relationship between web hits and references 20 years later. This is a key reason why the overall time structure of references is worth looking at to help review the current fashion show. The distribution as a whole may indicate the sorting out speed at which much of research activity is either abandoned or integrated into other work. In particular the tail of the distribution may be valuable in the identification of lasting research influences.

APPENDIX

This paper covers a period involving an enormous change in technology in publication. Within a few years enormous, more or less complete, data bases will have been created and statistical analysis of many aspects of web-browsing and the utilization of electronically available documents will be feasible. Before we are in a position to evaluate the meaning of "hit counts" on scholarship considerable attention must be paid to conceptual problems. We do not attempt this here. However we close with two graphs provided by our web controller, Glena Ames of recent downloads of 925 CFPs and 353 CFDPs.

Cowles Foundation Papers Downloaded
November 2001-December 2003
(925 papers on-line)



Cowles Foundation Discussion Papers Downloaded
November 2001-December 2002
(353 papers on-line)