SCIENCE AND TECHNOLOGY CITATION ANALYSIS: IS CITATION NORMALIZATION REALISTIC?

BY

Dr. Ronald N. Kostoff
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

Dr. Wendy L. Martinez
Office of Naval Research
800 N. Quincy St.
Arlington, VA 22217

ABSTRACT

One method for assessing quality of research outputs across different technical disciplines is comparing citations received by the research output documents. However, cross-discipline citation comparison studies require discipline normalization, in order to eliminate discipline differences in cultural citation practices and discipline differences in number of active researchers available to cite. The 'definition' of, and number of documents used to represent, a discipline become critical. This study attempted to determine whether the citation characteristics (average, median) of a discipline's domain stabilized as the domain's size was decreased. A sample of papers (classified as *research articles only*, not review articles, by the Institute for Scientific Information) published in the journal Oncogene in 1999 was clustered hierarchically, and the citation averages and medians were computed for each cluster at different cluster hierarchical levels. The citation characteristics became increasingly stratified as the clusters were reduced in size, raising serious questions about the credibility of a selected denominator for normalization studies. An interesting side result occurred when all the retrieved articles were sorted by number of citations. Thirteen of the fifty most highly cited *research articles* had 100 or more references, whereas zero of the fifty least cited *research articles* had 100 or more references.

# Report Documentation Page	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE **08 SEP 2004**	2. REPORT TYPE	3. DATES COVERED **-**

4. TITLE AND SUBTITLE **SCIENCE AND TECHNOLOGY CITATION ANALYSIS IS CITATION NORMALIZATION REALISTIC**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) **Ronald Kostoff; Wendy Martinez**	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Office of Naval Research,800 N. Quincy St.,Arlington,VA,22217**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

One method for assessing quality of research outputs across different technical disciplines is comparing citations received by the research output documents. However, cross-discipline citation comparison studies require discipline normalization, in order to eliminate discipline differences in cultural citation practices and discipline differences in number of active researchers available to cite. The ‡definitionŸ of, and number of documents used to represent, a discipline become critical. This study attempted to determine whether the citation characteristics (average, median) of a disciplineŸs domain stabilized as the domainŸs size was decreased. A sample of papers (classified as research articles only, not review articles, by the Institute for Scientific Information) published in the journal Oncogene in 1999 was clustered hierarchically, and the citation averages and medians were computed for each cluster at different cluster hierarchical levels. The citation characteristics became increasingly stratified as the clusters were reduced in size, raising serious questions about the credibility of a selected denominator for normalization studies. An interesting side result occurred when all the retrieved articles were sorted by number of citations. Thirteen of the fifty most highly cited research articles had 100 or more references, whereas zero of the fifty least cited research articles had 100 or more references.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES **36**	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**			

INTRODUCTION

Citation analysis is the quantitative and qualitative analysis of references in published documents (Narin, 1976; Kostoff, 2001). It is used mainly to identify historical trends in research disciplines, identify seminal documents, identify citer characteristics, and evaluate researcher/ research organization impact. Number of citations received by a document is a function of many variables, two of the most prominent being quality of the document's contents and number of researchers in the discipline(s) addressed by the document. To factor out the discipline effect (researcher candidate pool), especially when comparing research units across disciplines, some type of normalization is required. Various types of normalization have been used, including discipline normalization and journal normalization (Schubert and Braun, 1996). All these methods are founded on the belief that a discipline with nominal citation characteristics can be defined, thereby allowing some type of credible normalization.

The purpose of the present article is to examine citations of published papers in a given domain, allow the domain to get smaller, and ascertain whether isocitation regions of documents become relatively size-independent (the region-average citations would remain approximately constant as the region size changes). The approach started with a collection of documents from a technical 'discipline', performed document clustering that grouped the documents by similarity, allowed the groupings to get smaller, and thereby allowed the constituent documents of each group to become more similar in technical content. If the average group member citation value changed with size, this would raise questions as to whether any of the groups could be used as a denominator for clustering, and would raise more serious questions about whether credible normalization is possible.

METHODOLOGY AND RESULTS

Toward that end, we selected a discipline-focused journal (Oncogene), and downloaded 490 records (with Abstracts) for 1999, from the Science Citation Index (SCI). Each record was classified by the SCI as *a research article*; none were classified as review papers or otherwise. For each record, we tabulated #references, #citations, #keywords, #Abstract words, and #title words.

We examined the relationships among #Abstract words, #cites, and #refs. We first sorted based on #Abstract words, but found no significant relationship of #cites with # Abstract words. Both the top 50 and the bottom 50 records had twelve articles with 40 or more cites. However, the top 50 had zero articles with more than 100 references, whereas the bottom 50 had seven. We then sorted by #cites. Thirteen of the top fifty had 100 or more references, whereas zero of the bottom 50 had 100 or more references.

We then used our document partitional clustering algorithm (CLUTO) to generate a four level hierarchical tree (taxonomy) structure (Karypis, 2004; Zhao, 2004) from the papers' Abstracts. Most of CLUTO's clustering algorithms treat the clustering problem as an optimization process that seeks to maximize or minimize a particular clustering criterion function defined either globally or locally over the entire clustering solution space. CLUTO uses a randomized incremental optimization algorithm that is greedy in nature, and has low computational requirements.

For the first hierarchical level, the clustering algorithm split the total database into two categories. As shown in Table 1, for average cites, one of the clusters had an average document citation of 27.4 citations per document, and the other had an average citation of 27.3. For the second level, the algorithm split each first level category into two sub-categories, so that we had four second level categories. For the third level, the algorithm split each second level category into two categories, and for the fourth level, the algorithm split each third level category into two sub-categories. The lowest (fourth level) clusters averaged thirty papers each. Then, for each category in each level, we computed both the average and median number of citations.

We found that as the domains became smaller and more focused, and the Abstracts in each domain (cluster) became more similar in technical content, the average and median citations became more stratified (see Table 1). This suggests that a different method for computing citation normalization factor is required than presently used. While our demo was performed on the papers in a single journal, we wouldn't have to limit the source to a single journal in practice. We could use a query-based retrieval, and cluster the retrieved articles thematically. The key point is to arrive at thematically very similar articles in each cluster to be used as a basis for comparison.

TABLE 1

AVERAGE CITES (STANDARD DEV) TOTAL # PAPERS			
LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4
		22.84615 (17.85385) 52	20.25 (14.61734) 16
	29.45333 (47.80168) 150		24 (19.19523) 36
		32.95918 (57.50247) 98	32.2 (65.26368) 60
27.40351 (40.46126) 228			34.15789 (43.29129) 38
		19.825 (14.25030) 40	23.08696 (16.07910) 23
	23.46154 (19.51269) 78		15.41176 (10.17385) 17
		27.28947 (23.43006) 38	31.52632 (28.88746) 19
			23.05263 (16.00164) 19
		30.93902 (39.50569) 82	29.46875 (20.18300) 37
	27.98658 (34.06769) 149		31.88 (48.16537) 50
		24.37313 (25.75045) 67	23.72727 (24.57675) 33
27.27099 (33.17963) 262			25 (27.19625) 34
		22.62687 (24.02450) 67	23.41176 (30.88896) 34
	26.32743 (32.09707) 113		21.81818 (14.32317) 33
		31.71739 (40.83498) 46	25.76471(38.95434) 17
			35.2069 (42.17428) 29

MEDIAN CITES (Inner Quartile Range) TOTAL # PAPERS			
LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4
		16 (17.85) 52	21 (14.62) 16
	18 (47.80) 150		16 (19.20) 36
		18 (57.50) 98	17 (65.26) 60
18 (40.46) 228			26 (43.29) 38
		16 (14.25) 40	19 (16.08) 23
	20 (19.51) 78		12 (10.17) 17
		24 (23.43) 38	28 (28.89) 19
			22 (16.00) 19
		24 (39.51) 82	24 (20.18) 37
	19 (34.07) 149		24 (48.17) 50
		17 (25.75) 67	17 (24.58) 33
19 (33.18) 262			17 (27.20) 34
		15 (24.02) 67	14 (30.89) 34
	18 (32.10) 113		22 (14.32) 33
		21 (40.84) 46	11 (38.95) 17
			28 (42.17) 29

We then examined those articles (records) with 100 or more references, and evaluated their citation ranking in their level 4 (lowest) category. The results are shown in Table 2 below, ordered by category number.

TABLE 2

CITATION RANK IN
TAXONOMY LEVEL 4
ARTICLES WITH 100 OR
MORE REFS

CATEG#	#REFS	#CITES	RANK
3	345	471	1/60

3	111	154	2/60
4	128	232	1/38
4	137	22	20/38
5	176	50	2/23
5	101	17	13/23
7	165	133	1/19
7	187	65	2/19
7	136	31	7/19
8	141	55	1/19
9	108	19	24/32
10	213	318	1/50
10	187	56	4/50
11	157	123	1/33
11	119	56	3/33
12	106	139	1/34
12	139	39	5/34
12	127	23	8/34
15	188	162	1/17

The first row can be interpreted as follows. In the first category that had an article with over 100 references, category 3 of level 4, this article had 345 references and 471 citations, and it ranked first (out of 60 records in that category) in citations in that category. Thus, out of the 19 records in the table, 8 records were first in their respective level 4 categories, 3 were second, and 1 was third.

If we raise the threshold on cutoff to 150, or even 200 references, the results are even more striking. There are eight records with 150 or more references, of which five rank first in their respective categories, two rank second, and one ranks fourth. There are two records with 200 or more references, and both rank first in citations in their relatively large categories.

Thus, the articles that have large numbers of references tend to be highly cited, especially when compared to strongly thematically related articles.

We then examined the other end of the spectrum. Table 3 shows the metrics for articles that contained the least references. There were 15 records with 18 or less references. Three were last in their respective categories in citation ranking, and nine were in the bottom half. However, three were in the top quarter.

TABLE 3

ARTICLES WITH 18 OR LESS REFS

CATEG#	#REFS	#CITES	RANK
1	17	6	16/16
4	16	34	13/38

4	11	13	27/38
6	15	26	3/17
6	17	18	5/17
7	16	35	5/19
9	9	6	28/32
9	14	2	32/32
12	16	9	29/34
12	16	27	8/34
14	16	52	1/33
14	17	23	15/33
14	16	11	22/33
16	18	25	16/29
16	18	4	29/29

Finally, we examined the characteristics of the 16 articles that ranked at the top of their respective categories in terms of citations, and the 16 articles that ranked at the bottom. The next two tables, 4 and 5, display the metrics.

TABLE 4

HIGHEST CITED RECORDS IN EACH CATEGORY - LEVEL 4

#REFS	#ABSWD	#CITES	#TTLWD	#KEYWD	CLUST#	ORDER-
72	112	243	8	25	49	63
106	117	139	19	25	11	50
213	136	318	8	23	55	39
345	139	471	15	20	34	13
38	141	67	16	21	62	23
188	157	162	33	24	0	61
16	158	52	9	10	36	58
157	164	123	16	23	28	44
141	165	55	21	18	25	30
34	172	42	14	20	42	25
39	189	148	9	17	57	54
128	214	232	17	27	4	19
55	228	85	8	20	45	34
165	240	133	9	23	18	27
54	261	81	20	19	20	4
72	283	45	25	22	16	2
113.9375	179.75	149.75	15.4375	21.0625	<<<<<<<<<	AVERAGES OF ABOVE
89	164.5	128	15.5	21.5	<<<<<<<<<	MEDIANS OF ABOVE

TABLE 5

LOWEST CITED RECORDS IN EACH CATEGORY - LEVEL 4

#REFS	#ABSWD	#CITES	#TTLWD	#KEYWD	CLUST#	ORDER-
24	148	0	14	19	16	2
29	105	4	23	15	17	5
20	172	1	17	25	13	10
29	189	0	8	21	24	18
29	235	2	20	21	58	24
24	191	4	12	20	42	25
28	189	4	13	18	27	29
50	195	4	9	18	9	32
14	185	2	20	17	41	36
38	179	0	19	19	59	40
32	305	5	15	19	51	43
43	217	7	16	22	37	49
65	189	2	9	23	60	51
54	184	3	10	21	44	55
52	137	0	22	21	0	61
18	136	4	10	14	54	64
34.3125	184.75	2.625	14.8125	19.5625	<<<<<<<<<	AVERAGES OF ABOVE
29	187	2.5	14.5	19.5	<<<<<<<<<	MEDIANS OF ABOVE

The major difference in both the average and median values is number of references.

SUMMARY AND CONCLUSIONS

In summary, to compare the quality/ impact of different research papers as represented by citations, the papers should be as similar thematically and typically (research article, review article, etc) as possible. Publication dates, journals, and other factors should be normalized, where possible. For the Oncogene test case, segregation according to thematic similarity resulted in changing group citation averages. This suggests that a meaningful 'discipline' citation average may not exist, and the mainstream large-scale mass production semi-automated citation analysis comparisons may provide questionable results. It further suggests that meaningful cross-discipline citation comparisons require the manually intensive approach of identifying those few research papers most closely related to the paper of interest, and normalizing on those papers (Appendix; Kostoff, 2002). Finally, it confirms what many research evaluators recognize instinctively: there are really relatively few very thematically similar technical articles in any discipline, and any metrics used to evaluate research should be based on this reality.

MAIN TEXT REFERENCES

Karypis G. CLUTO—A Clustering Toolkit. http://www.cs.umn.edu/˜cluto

Kostoff RN, Del Rio JA, García EO, Ramírez AM, Humenik JA. (2001). Citation Mining: Integrating Text Mining and Bibliometrics for Research User Profiling. Journal of the American Society for Information Science and Technology. 52:13. 1148-1156.

Kostoff, R. N. (2002). Citation Analysis for Research Performer Quality. Scientometrics. 53:1. 49-71.

Narin F. (1976). Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity (monograph). NSF C-637. National Science Foundation. Contract NSF C-627. NTIS Accession No. PB252339/AS.

Schubert, A., Braun, T. (1996). Cross-Field Normalization of Scientometric Indicators, Scientometrics. 36:3. 311-324.

Zhao Y, Karypis G. Criterion Functions for Document Clustering: Experiments and Analysis. Machine Learning, in press.

APPENDIX – CITATION ANALYSIS OF RESEARCH PERFORMER QUALITY

BACKGROUND: Citation analysis for evaluative purposes typically requires normalization against some control group of similar papers. Selection of this control group is an open question.

OBJECTIVES: Gain a better understanding of control group requirements for credible normalization.

APPROACH: Performed citation analysis on prior publications of two proposing research units, to help estimate team research quality. Compared citations of each unit's publications to citations received by thematically and temporally similar papers.

RESULTS: Identification of thematically similar papers was very complex and labor intensive, even with relatively few control papers selected.

CONCLUSIONS: A credible citation analysis for determining performer or team quality should have the following components:

*Multiple technical experts to average out individual bias and subjectivity

*A process for comparing performer or team output papers with a normalization base of similar papers
*A process for retrieving a substantial fraction of candidate normalization base papers
*Manual evaluation of many candidate normalization base papers to obtain high thematic similarity and statistical representation

I. INTRODUCTION

In the evaluation of science and technology (S&T), whether ongoing or proposed programs, a key criterion is the track record of the proposer or performer. Past analyses [DOE, 1982; Kostoff, 1997a] have shown that, typically, the criterion of Team Quality is the major determinant of program or project quality. Many qualitative and quantitative approaches have been used for the purpose of determining Team Quality [Kostoff, 1997a]. None are viewed as adequate in a stand-alone mode, and present practice is to use multiple approaches to determine Team Quality [Martin, 1983; Kostoff, 1997b].

One of the more widely used of these approaches, especially applicable to research, is citation analysis. For proposer quality assessment, citation analysis consists of counting citations to documents produced by the proposer's research unit, then comparing this citation count to numbers of citations received by similar documents from other research units. The assumption is then made that documents with higher relative numbers of citation counts have more impact than those with lower citation counts, and are of higher quality from the citation metric perspective.

While this approach appears rather straight-forward and deceptively simple, it is intrinsically very complex. This appendix will illuminate the complexities, and show that high quality S&T citation analysis requires technical experts performing very manually intensive comparisons with very subjective judgements. It will show further that the automated assembly-line approaches to citation analysis, widely used by the decision aid community today, are highly uncertain at low-to-mid citation levels characteristic of most research.

After a background description of the problem, the analytical techniques developed for the citation analysis will be presented. Two illustrative examples of the use of citation analysis to support proposal review will be presented. Because of the confidentiality agreements operable for proposal review, all information that identifies either the proposing organization or the potential science and technology sponsor will be removed. The results of the analysis will then be presented, followed by summary and conclusions that emphasize the lessons learned from using these techniques. Special emphasis will be placed on requirements for thematic similarity between the target documents and the external documents against which they are compared.

III. BACKGROUND

In the present context, citation is referencing, in a document, the work of another individual or group. The work referenced can exist in many forms, although the most common use is reference of another document. Citation analysis is the examination of the multiple dimensions and myriad facets of citations for the purpose of understanding the many impacts of the target documents of interest.

Citation counts resulting from citation analyses are usually classified as outputs, but they are neither outputs nor outcomes. While they are closer to outputs than outcomes, since they can be used in relatively short range analyses and they do not impact the larger problems characteristic of outcomes, they are not under the direct control of the performer. More correctly, they are near-term impacts.

Modern day interest in studying and developing the citation process accelerated after WW2 [e.g., Zachlin, 1948, Zirkle, 1954]. However, the origins of citation analysis as a widespread bibliometrics tool can be traced to the mid-1950s, with Garfield's proposal for creating a citation index [Garfield, 1955]. As the Science Citation Index (SCI) was developed, along with companion citation indices, the computer revolution and associated information technology developed in parallel. The combination of SCI, massive information storage, and rapid information retrieval laid the foundation for a multi-application S&T evaluation capability.

The foundations of modern traditional citation analysis were established by Garfield [1955, 1963, 1964, 1965, 1966, 1970] and CHI, Inc [Narin, 1975, 1976, 1984, 1994, 1996; Albert, 1991], and extended to co-citation analysis by Small [1973, 1974, 1977, 1981, 1985], Sullivan [1977, 1979, 1980], and Marshakova [1973, 1981, 1988].. The practice of citation analysis has been extended further by groups at the Hungarian Library of Sciences [Schubert, 1986, 1993, 1996; Zsindely, 1982] and the University at Leiden [Moed, 1986; Nederhof, 1987; Braam, 1988, 1991; VanRaan, 1991, 1993, 1996; Davidse, 1997]. A broad summary of the status of citation analysis is contained in a recent festschrift to Eugene Garfield [Festschrift, 2000].

Traditional citation analysis is presently used both at the micro and macro scales. It is used at the micro level, especially in academia, to evaluate components of impact of a given published document, or the documents published by a given researcher or research group. It is used at the macro level to evaluate technical discipline or national outputs. Because of the large numbers of documents and subsequent citations that exist in macro level analyses, semi-automated techniques have been developed to handle the data efficiently. As time has proceeded, these semi-automated techniques have diffused toward micro level application.

Citation analysis has two components. The first component is counting of citations to a document or group of documents, depending on the purpose of the analysis. The second component is placing these citation counts in a larger context through a comparison and normalization process, to provide meaning to the numbers of counts obtained.

Many articles have been written about problems inherent in the traditional citation analysis process [e.g., Geisler, 2000; MacRoberts, 1989, 1996; Kostoff, 1998]. There are two main categories of problems: those associated with the counts of citations, and those associated with the comparisons of counts of citations. The problems associated with counts of citations can be sub-divided further into problems associated with the quantity of the underlying data, and problems associated with the quality of the underlying data.

III-A. Problems with Citation Counts

III-A-1. Problems with Quantity of Underlying Data

The main resource available for performing citation analysis today is the SCI. The number of candidate articles to be used in a citation analysis is limited to the number of articles in the total SCI. This total is limited by the following sequence of steps.

a) There is approximately $500 billion-$800 billion/ year worth of S&T being performed globally today, depending on one's definition of S&T. Only a small fraction of the S&T performed is documented. While there are many reasons for this [Kostoff, 2000a], basically there are more disincentives to publishing than incentives.

b) Of the S&T performed that eventually gets documented, only a very modest fraction is accessed by the SCI (or any single database). There are tens of thousands each of internal and external technical reports, classified reports and papers, workshop and conference proceedings, journals, magazines, newspapers, and patents resulting from the S&T performed and published annually. Yet, the SCI accesses only about 5600 journals presently. While these accessed journals tend to be the highest quality peer-reviewed research journals, they represent only a fraction of S&T that is documented.

c) Of the documented S&T that is accessed by the SCI, only a fraction reaches the average analyst performing citation analysis. The main reason is the extremely poor information retrieval techniques actually used by the technical community [Kostoff, 2000b].

Thus, the citation counts derived from the records in the SCI under-represent the total referencing of prior work by the global technical community, and there is no evidence that this under-representation is homogeneous across disciplines or sub-disciplines.

III-A-2. Problems with Quality of Underlying Data

The problems with citation data quality translate into problems with the citation selection process (i.e., the approach used by authors to select references for inclusion in their papers). The issues related to the sociological and cultural aspects of how people cite have been raised by the references cited above, and will not be repeated here. Suffice it to say that the combination of quantity and quality problems with citations places strong limits on the degree to which citations can be used as a stand-alone metric. This is especially true for documents that receive mid and low level numbers of citations (i.e., the vast majority of documents published); the very highly cited documents (a very small fraction of all articles published) are in a class by themselves, and modest margins of error in interpreting their citation counts don't affect overall conclusions about their impact.

III-B. Problems with Citation Comparisons

Problems with citation count comparisons form the focus of this paper. Whether applied to micro or macro scale problems, citation count comparisons have received insufficient attention, and offer further severe constraints on the credibility of present day citation analyses. There are two main types of potential citation count comparisons: comparison of counts to an absolute standard, and comparison of counts to a relative standard. The former comparison is analogous, in the physical sciences, to comparing actual engine efficiencies to maximum engine efficiencies possible (Carnot efficiencies). The latter comparison is analogous to an athletic competition, where one group's performance is compared to another group's performance. One problem with the latter comparison is that the performance of a group is never related to its potential, only to the performance of another 'similar' group. The latter comparison is used in essentially all citation analyses today. This issue of comparison with absolute or relative standards was examined in a 1997 paper [Kostoff, 1997c], and will not be addressed further.

Citation count comparisons are necessary because of the high variability of citation counts with different parameters. Citation counts depend strongly on the specific technical discipline, or sub-discipline, being examined. The funding and number of active researchers can vary strongly by sub-discipline, and these numbers of researchers affect the numbers of citations directly. The maturity of the sub-discipline affects the numbers of citations, since the basic research community is oriented more toward publishing than the applied research or technology development communities. The breadth of the sub-discipline can affect citation counts, since more focused disciplines will concentrate citations into fewer key researchers. The classification and proprietary levels can vary sharply by sub-discipline, and can strongly affect what gets published and therefore cited in open-literature publications. The documentation and citation culture can vary strongly by sub-discipline. Since citation counts can vary sharply across sub-disciplines, absolute counts have little meaning, especially in the absence of absolute citation count performance standards.

Thus, in order to provide meaning and context to citation counts for performance evaluation in traditional citation analysis, some type of citation count normalization is required. The main normalization approaches used in traditional citation analyses are described in an excellent review article [Schubert, 1996]. They can be summarized as follows:

1) Reference standards based on prior sub-field classification

Journals are classified into a number of science sub-fields. Since some journals are single discipline, and some multi-discipline, percentage weights are assigned to each journal indicating their connection with the different sub-fields. According to Schubert [1996], the method works only at a higher (macro) statistical level; i.e., if the sample under study is large and mixed enough to support the validity of such a statistical approach. Further according to Schubert [1996], for micro level analyses, it is sometimes unavoidable to use a classification scheme concerning not only the journals but every single paper. Schubert proceeds to point out that such classification schemes are enclosed in some specialized databases, such as in the *Physics Briefs*, to classify each paper into one or more of ten first-level and many lower-level sub-fields of physics.

2) Journals as reference standards

Primary journals in science are generally agreed to contain coherent sets of papers both in topics and professional standards. According to Schubert [1996], it seems justified to regard the set of regular authors of a journal as reference standard for any single author (or team of authors), the set of institutions regularly publishing in the journals as reference standard of any single institution, the citation rate of the set of papers published in the journal (or of a properly selected subset) as reference standard of any single paper. Also according to Schubert [1996], one may thus expect that any difference in productivity, citation rate or other scientometric indicators reflects differences in inherent qualities.

3) Related records as reference standards

Subject matter similarity between two documents is measured by the number of shared references. According to Schubert [1996], bibliographic coupling appears to be one of the most selective and flexible techniques of reference standard selection, but “because of its high requirements in time and effort, its use can be suggested only in micro or meso-level”.

It is the present first author’s contention that none of the above normalization methods are adequate for precise normalization, since they do not provide sufficient resolution for distinguishing among the lower level sub-fields. Inability to distinguish precisely among sub-fields translates, in some cases, to substitution of far different magnitude numbers for the normalization base. The next section will show some of the effort required for more precise normalization comparisons.

IV. ANALYSIS TECHNIQUES AND ISSUES

IV-A. First proposal

A few years ago, the first author was asked, by a potential sponsor, to evaluate an S&T proposal generated by organization XXXX. While there were a number of criteria that had to be evaluated relative to technical quality and relevance of the proposal to the potential sponsor’s mission, one key criterion was the quality of the proposer’s research team. It was decided to evaluate team quality through evaluation of the research team’s various outputs and outcomes, using citation analysis and other metrics. This section focuses on the citation analysis component used..

The proposal and accompanying material presented many different types of outputs from XXXX researchers. Assessing the quality and impact of those outputs was complex, especially since they covered more than one research area. The following procedure was used as a first-order estimate of quality/ near-term impact of XXXX’s output, and thereby of the research team.

The citations of selected XXXX publications were compared against those of thematically similar non-XXXX publications (a control group of publications), using a pair-wise comparison approach. Specifically, all XXXX publications for 1996 (38 documents), as identified in the Web version of the Science Citation Index (SCI), were compared with thematically similar non-XXXX publications from the SCI.

[1996 was selected as a compromise year. The first author wanted to examine recent documents that reflected current management and staff of XXXX, but also wanted to insure that sufficient time had passed since publication such that citations had a reasonable chance to accumulate. Figures 1 and 2, titled Citing Papers Time Distribution, show the yearly and cumulative numbers of citing papers as a function of time, for 1996 and

1993, respectively. For 1996, the citing papers (for all the XXXX papers published in 1996) show a linearly increasing cumulative trend up to and including 2000. For 1993, the citing papers (for all the XXXX papers published in 1993) show more of an S-curve trend. While 1993 shows a leveling off of the citations, and would therefore have been a better year to select from that perspective, it was judged to be too far in the past to be relevant for assessing the quality of present XXXX staff and management. Citations from 1996 should almost be ready to level off, if the 1993 distributions can be extrapolated to 1996, and therefore 1996 was selected.]

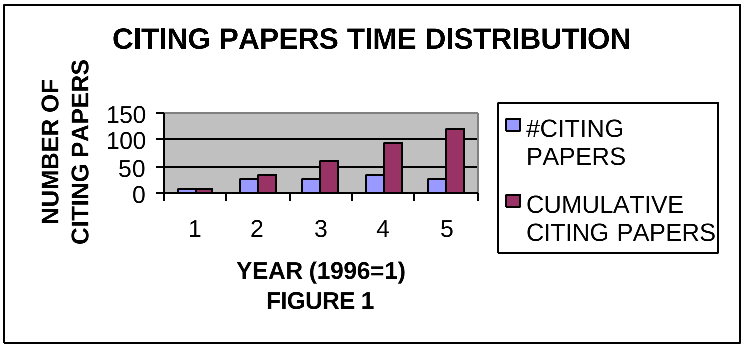

FIGURE 1

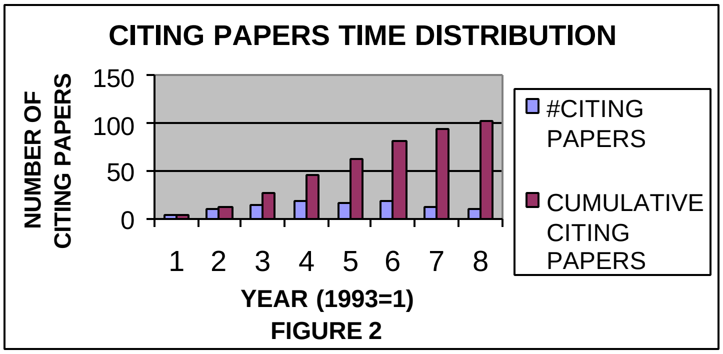

FIGURE 2

Ideally, the size of the control group for each paper should be statistically representative of the total thematically similar non-XXXX papers in the SCI, since the purpose of the citation analysis is to compare the citation performance of each proposer's paper to the aggregate of the relevant performer community.. Practically, resource and time constraints placed severe linits on the size of the control group. Specifically, for each of the 38 papers published in 1996 (hereafter referred to as the target papers), three non-XXXX papers thematically and temporally similar to the target papers were selected. If 1996 papers with the requisite thematic characteristics could be identified, they were given first priority in the selection, to insure temporal normalization. If 1996 papers could not be identified, then 1997 papers were selected. Thus, the results are conservative with respect to XXXX.

Selection of papers in the SCI thematically similar to the target paper depends strongly on the study's purpose and objectives, the mission of the performing organization, the degree of focus of the paper's theme, the size of the research paper pool from which to choose, and the level of technical description in the paper's SCI Abstract. The relation to study purpose is especially important, and is often overlooked. Specifically, is the purpose of the study to evaluate the 'job right' quality of the performer (i.e., is the specific task selected being performed with the latest tools and techniques to achieve the specific objectives?), or is the purpose of the study to evaluate the 'right job' quality of the performer (i.e., have the right task and right objectives been selected?). If the focus is on 'job right' quality, then the thematically similar papers will be limited to a very narrow area of inquiry. If the focus is on 'right job' quality, then the focus of thematically related papers can be expanded greatly.

For example, suppose that a researcher being evaluated was performing acoustic studies in the 100 KHZ small object detection regime. If the performing organization's mission in acoustics was limited to performing studies only in this regime, and if the quality determination was phrased as how well the researcher was performing relative to other researchers studying the 100 KHZ regime, then the thematically similar papers would all be focused narrowly around frequencies of 100 KHZ. The study reduces to determining the most cited papers at 100 KHZ. If, however, the organization's mission in acoustics provided flexibility in selecting the frequency regime to study, and the organization *chose* to focus on the 100 KHZ regime, then thematically related papers could include those in a broader range of frequency regimes. The study reduces to determining the most cited paper in mid-high frequency acoustics. The choice of journal as reference standard, described previously and referenced in Schubert [1996], relates strongly to the latter definition of organization mission, where essentially any paper in an acoustics specialty journal could serve as a reference standard. The practical implications of 'job right' vs 'right job' comparisons are that papers with substantially higher citation counts could be included in the normalization pool as the allowed definition of thematic similarity becomes broadened.

Selection of papers thematically similar to the target paper was very difficult, time-consuming, and subjective. This was especially true for the broad-based analyses. The selection was more straightforward for the much more limited specific technology papers, since these more focused areas seemed to have many researchers working related problems. The author believes that the subjectivity involved in selecting thematically similar papers is a major source of uncertainty of the results. A rigorous study, in addition to having the rigorous information retrieval and statistical sampling processes mentioned in the next two paragraphs, requires the use of multiple evaluators for the same target papers to average out evaluator subjective bias.

Many of the applied research papers combined analytical technique advancement with novel application advancement. It was not always possible to have thematic similarity for both technique and application, especially in those research areas with relatively few performers, and typically a choice had to be made between technique and application for determining thematic similarity.

Two important issues were i) determining the number of thematically similar candidate papers in the pool from which to choose, and then ii) determining the number of papers to select from the pool. First, in a rigorous study, candidate thematically similar papers would be identified by the most rigorous processes available. In the first author’s information retrieval studies [Kostoff, 1997d, 2000b], a manually intensive iterative approach using computational linguistics and bibliometrics is used to identify the full scope of relevant literature papers for each specific topic studied. For the present study, this would have required 38 such literature searches. In the time available, even one such rigorous literature search was not feasible. A very approximate approach was used.

Second, the number of papers to select from the candidate pool should have the greatest thematic similarity, and be representative statistically. Again, this would have required poring over hundreds, or thousands, of similar papers, and selecting a substantial number of the most representative thematically. Again, a small sampling approach was used because of time exigencies.

The first selection step was to examine the Related Records field of the SCI for a given target paper. This field contains papers that have at least one reference in common with the target paper, as stated previously [Schubert, 1996]. Papers that share references tend to be similar thematically, but this is not always true, and the relation between thematic similarity and number of shared references is not always monotonic.

Because of time constraints, a limited number (three) of thematically related papers was examined for each target paper. If three records thematically similar to the target paper could be identified from the Related Records papers, the selection was completed for that target paper. If three records could not be identified, then key words from the target paper’s Abstract/ Title/ Keyword fields were used to search the SCI for related records. This approach was substantially more time consuming than the already time-consuming Related Records approach.

FIGURE 3 - CITATION AND FIGURE OF MERIT DATA

A	B	C	D	E	F	G	H	I	J	K	L
REC#	PAP	SELF	PAP1	PAP2	PAP3	AVER		MED		STD	
	CIT	CIT	CIT	CIT	CIT	CIT	FOM1	CITE	FOM2	DEV	FOM3
								S			
										CIT	

1	4	1	3	3	23	9.667	0.293	3	0.571	11.55	-0.49
2	2	1	9	7	21	12.33	0.14	9	0.182	7.572	-1.36
3	0										
4	0		5	1	2	2.667	0	2	0	2.082	-1.28
5	0		5	6	9	6.667	0	6	0	2.082	-3.2
6	3	2	3	4	4	3.667	0.45	4	0.429	0.577	-1.15
7	0		11	14	4	9.667	0	11	0	5.132	-1.88
8	1	1	1	3	2	2	0.333	2	0.333	1	-1
9	6	3	3	7	5	5	0.545	5	0.545	2	0.5
10	5	0	2	5	16	7.667	0.395	5	0.5	7.371	-0.36
11	5	3	5	2	14	7	0.417	5	0.5	6.245	-0.32
12	2	2	3	3	2	2.667	0.429	3	0.4	0.577	-1.15
13	1	0	4	4	5	4.333	0.188	4	0.2	0.577	-5.77
14	5	2	6	4	9	6.333	0.441	6	0.455	2.517	-0.53
15	7	4	15	5	12	10.67	0.396	12	0.368	5.132	-0.71
16	5	5	3	7	1	3.667	0.577	3	0.625	3.055	0.436
17	4	4	8	4	6	6	0.4	6	0.4	2	-1
18	9	4	38	2	13	17.67	0.338	13	0.409	18.45	-0.47
19	4	2	3	7	7	5.667	0.414	7	0.364	2.309	-0.72
20	2	1	2	6	8	5.333	0.273	6	0.25	3.055	-1.09
21	0	0	2	5	16	7.667	0	5	0	7.371	-1.04
22	1	1	13	8	9	10	0.091	9	0.1	2.646	-3.4
23	24	20	5	2	7	4.667	0.837	5	0.828	2.517	7.682
24	4	0	4	22	8	11.33	0.261	8	0.333	9.452	-0.78
25	0										
26	0										
27	3	0	11	14	2	9	0.25	11	0.214	6.245	-0.96
28	2	2	3	3	4	3.333	0.375	3	0.4	0.577	-2.31
29	4	4	8	10	6	8	0.333	8	0.333	2	-2
30	2	2	3	3	13	6.333	0.24	3	0.4	5.774	-0.75
31	1	1	2	4	5	3.667	0.214	4	0.2	1.528	-1.75
32	0										
33	6	6	13	26	3	14	0.3	13	0.316	11.53	-0.69
34	0	2	2	4		3	0	3	0	1.414	-2.12
35	3	1	2	5	16	7.667	0.281	5	0.375	7.371	-0.63
36	0		2	7	1	3.333	0	2	0	3.215	-1.04
37	2	1	5	22	4	10.33	0.162	5	0.286	10.12	-0.82
38	4	1	5	3	14	7.333	0.353	5	0.444	5.859	-0.57
SUM	115	74	197	200	252	AVER	0.297		0.324		-0.98

Once thematically similar records were identified, the citations for each of the four records were tabulated. Figures of merit were generated, and the citation performance of each target paper was compared with that of the three thematically related papers. The results are shown in Figure 3. Starting from the left, column A is the number of the record, column B is the citations of the target paper, column C is the self-citations of the target paper, columns D, E, F are the citations of the thematically similar papers (the Abstracts of papers 3, 25, 26, 32 did not contain sufficient information for similar papers to be identified), column G is the average citations of the thematically similar papers, column I is the median citations of the thematically similar papers, and column K is the standard deviation of the citations of the thematically similar papers. Columns H, J, L are figures of merit FOM1, FOM2, FOM3, respectively, defined as follows:

FOM1=citations of target paper/ (citations of target paper plus average citations of related papers)

FOM2=citations of target paper/ (citations of target paper plus median citations of related papers)

FOM3=(citations of target paper minus average citations of related papers)/ standard deviations of related papers.

FOM1 and FOM2 have the desirable properties of ranging between zero and unity, as well as equaling 0.5 when the target paper citations equal those of the average or median citations of the related papers. FOM3 removes the limitations of using absolute number values, and places the citation differences in the context of standard deviations.

This section ends with a note about the four papers that could not be evaluated due to insufficient information contained within the Abstract. Ideally, with unlimited time and resources, the full text target and control group papers would be read in their entirety. Practically, time is available for reading Abstracts only. Unfortunately, in the non-medical technical literature, and some of the medical literature, there are no requirements on the technical content of Abstracts. Consequently, many Abstracts contain very little technical detail, and they cannot be used in the citation process. This issue is addressed summarily in a letter to Science [Kostoff, 2001a], and in more detail in a letter to selected technical journal editors proposing the use of Structured Abstracts in all technical journals [Kostoff, 2001b].

IV-B. Second Proposal

In early 1998, the first author was asked to evaluate an S&T proposal for a different potential sponsor, generated by an organization (ZZZZ) different from the proposing organization (XXXX) of the first proposal. One critical component again was evaluation of team quality. This was a complex procedure for the second proposal, since most of the organization's publication outputs were co-authored with people from other organizations, and the author wanted to identify the quality of the contributions of researchers from organization ZZZZ only. Again, citation analysis was one of several methods used to gauge team quality, and this section reports on the citation analysis component only.

1. Database Examined and Process Used

One purpose of the study was to examine the citation impact on the technical community of the ZZZZ researchers who publish. Another purpose was to assess some estimate of the ZZZZ researchers' contribution to the published product. Two studies were performed. First, all the 1997 papers in the web version of the SCI that contained a ZZZZ author address were examined. The position of the ZZZZ author in the author list for each paper was highlighted. Citations for this group of papers were not examined, because of the recent date.

Second, all the 1993 papers that contained a ZZZZ author address were examined. 1993 was selected for two reasons. A four-year lag allows many (not all) citations to accumulate, and is sufficient to show differentiation in citation counts among papers. Also, 1993 was the third year that paper abstracts were included in the SCI, allowing more than title information to be obtained about a paper if necessary. Author position was highlighted again, and then the citations received by each paper with citations received by a non-ZZZZ authored paper of similar theme were compared.

V. RESULTS AND DISCUSSION

V-A. First Proposal

The results for the first proposal are as follows.

Figures 4 and 5, titled Citation Distribution Function, show the numbers of papers N(X) with X cites for 1993 and 1996, respectively. 63% of the 1993 target papers had either zero or one cites, and 37% of the 1996 target papers had either zero or one cites. For 1996, the average number of citations per target paper was three, of which 2/3 were self-cites. (No judgements are made about including or excluding self-cites. To make such judgements rationally, each full-text paper would have to be read, and the technical rationale for self-citation other than author self-gratification would have to made. Such a level of detail is beyond the scope of this study.) For 1993, the average number of citations per target paper was about 2.5. For 1996, the average number of citations per thematically related paper was about twice the number of target paper citations.

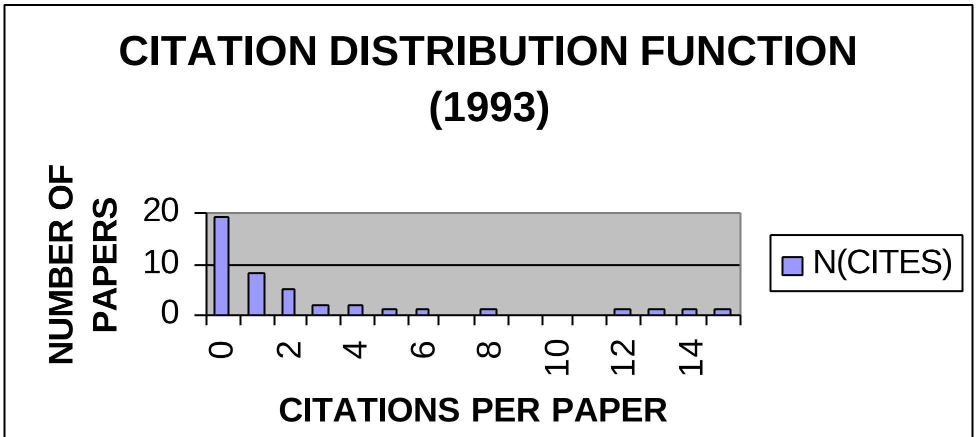

FIGURE 4

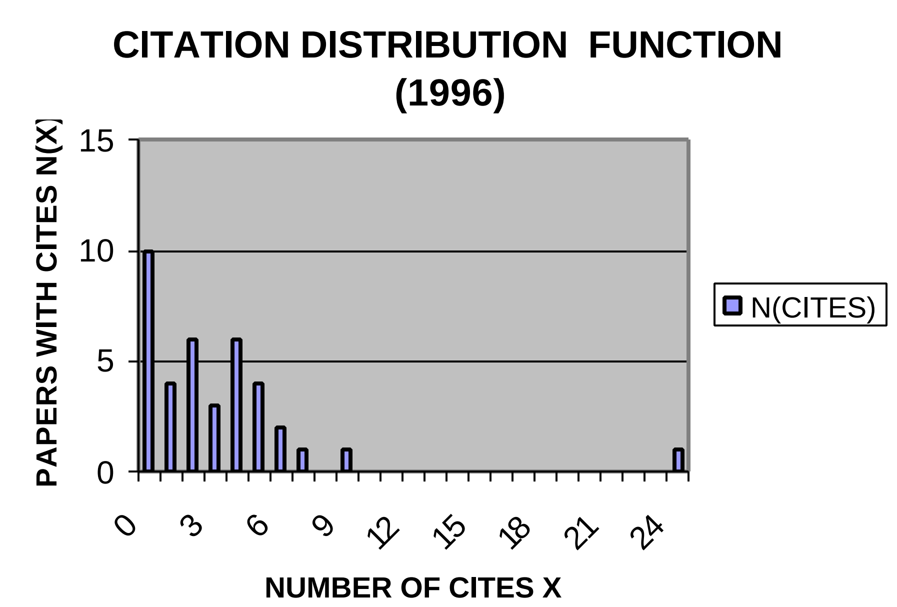

FIGURE 5

For 1996, the average value of FOM1 and FOM2 was about 0.3, and the average value of FOM3 was about minus one standard deviation. Thus, all three figures of merit gave essentially similar results. FOM1 and FOM2 were greater than 0.5 in less than ten percent of the target papers examined. In the best performing target paper, both in absolute citations and relative citations, 20 of the 24 citations were self-cites. This particular paper had many authors, and many of these authors cited the target paper in later publications.

Many of the research disciplines examined seem to have relatively few papers thematically related to the target paper. In addition, the absolute levels of citations are low, relative to other disciplines the author has examined. This suggests research into areas that have few performers, probably low funding, and therefore low citations.

V-B. Second Proposal

1. Results and Discussion

a. 1997 Database

In the 1997 database, there were 43 papers in the SCI with a ZZZZ address for the research unit. These papers had a total of 184 authors, with an average of 4.29 authors per paper, a median of 3 authors per paper, and a mode of 3 authors per paper. A Coefficient of Author Position (CAP) was defined as a measure of the ZZZZ author's location in the total author list. The definition of CAP was:

CAP=(x-1)/(n-1)

where x was the location of the ZZZZ author in the list, and n was the total number of authors in the list. Thus, if there were three authors in the list, and the ZZZZ author was third, CAP would equal one. If the ZZZZ author was first in this case, CAP would equal zero. If the paper had only one author, CAP was set equal to zero. Thus, the higher the value of CAP, the less was the relative contribution of the ZZZZ author.

There are two assumptions here. First, the ordinal positioning of any author in the list reflects his/ her relative contribution to the paper. In the absence of large power differential relationships (e.g., advisor/ student), this is probably a very reasonable assumption. In the presence of large power differential relationships, it may or may not be reasonable, but validation of the assumption would be next to impossible.

Second, the ordinal positioning can be quantified for computational purposes. There appears to be nothing in the literature that supports or rejects this assumption. For large numbers of papers undergoing citation analyses, anomolies will disappear, and quantification for estimation purposes may be reasonable. However, because of the uncertainty of the validity of this assumption, supplementary approaches were used to estimate the contribution of organization ZZZZ’s researchers to overall paper quality. In this particular case, there were no significant differences in final results among the different methods used.

The total value of CAP summed over the 43 papers was 26.27, with an average value of 0.61, a median value of .92, and a mode of 1. Most papers were multi-authored; there were only four papers with one author. To summarize these results, the preponderance of papers that include an ZZZZ research unit author address have multiple authors, and the ZZZZ author is usually at the end of this list. The typical paper in this database had about three authors, with the ZZZZ author being last.

b. 1993 Database

i. Author Position Study

In the 1993 database, there were 44 papers in the SCI with an ZZZZ address. These papers had a total of 126 authors, with an average of 2.86 authors per paper, a median of 3 authors per paper, and a mode of 3 authors per paper. The total value of CAP summed over the 44 papers was 18.97, with an average value of .43, a median double value of 0/.5 (half the papers had a CAP of zero, the other half had a CAP of .5 or greater) and a mode of 0. The typical paper in this database had about three authors, with the ZZZZ author being second.

In comparison with the 1997 database results, the total number of papers is about the same. The median and mode of authors per paper is the same, but the average has dropped by a third from 1997 papers to 1993 papers. More importantly, the average CAP value dropped by a third from 1997 to 1993, the median CAP value dropped by a half, and the mode plummeted from one to zero. Thus, in 1993, the ZZZZ authors were contributing significantly more to papers (as measured by their ordinal position in the authors list) than in 1997.

ii. Citation Comparison Study

For the 1993 database, citations of pairs of similar theme papers were compared. In particular, for a given paper with a ZZZZ author address in the list, a similar theme paper was selected from the Related Records field, and the number of citations received by each paper was transcribed and compared. The procedure used was to select the first 1993 paper from the Related Records field with a similar theme to the target paper (this procedure normalized publication date and theme), and compare each paper's citations. (In a very few cases, no 1993 papers could be found in the Related Records field, and a 1994 or 1992 paper of similar theme was used. In a very few cases, no similar theme paper could be found for 1992 or 1994.)

Then, the ratio of citations of the two papers was transcribed, and this ratio was placed in one of five bands: very high (VH), high (H), same (S), low (L), very low (VL).

'Very High', for example, meant that the ratio of citations received by the related paper to the citations received by the ZZZZ paper was very high, a subjective judgement made by observation. 'Same' meant that the numbers of citations received by the two papers were close, not necessarily identical. Typically, citations received by a few of the other related papers would be examined to ascertain the approximate range of citations, and then judgements about the significance of the differences in citation numbers would be made. Obviously, in a definitive or final study of this nature, there would need to be people involved who could judge if in fact themes were closely related, and there would need to be citation distribution studies of related papers to obtain a more quantitative basis for judging significance of differences.

The population of the five bands was as follows: 12(VH); 9(H); 14(S); 4(L); 1(VL), for a total of 40 pairs where the citations could be compared. While the mode is in the S band, the median is in the H band. Since half the papers in the database had a CAP of zero, all other things being equal one would expect six papers in the VH band to have a CAP of zero. In actuality, nine papers in the VH band had a CAP of zero. Thus, those papers with a VH figure of merit tended to have more ZZZZ lead authors than one would expect from the database overall average.

There were seven prolific ZZZZ authors, each of whom participated in three or more papers. The population of the five bands for these seven prolific authors was: 1(VH); 5(H); 9(S); 3(L); 0(VL). Compared to the overall 1993 database, where 52.5% of the ZZZZ papers were in the VH or H bands, these seven authors had 33% of papers in the VH and H bands. Also, for these seven authors, the average CAP was .6, the median CAP was 0.8, and the mode CAP was 1. For the 1993 database, the parallel numbers were .43 (av), 0/.5 (med), 0 (mode). Thus, while the more prolific authors had better relative citeability than the database average, these authors were closer to the end of the author listing than the database average.

iii. Discussion

The highlights of this author position study are:

* The preponderance of 1997 papers that include a ZZZZ author address have multiple authors, and the ZZZZ author is usually at the end of this list. The typical paper in this database had about three authors, with the ZZZZ author being last.

* In 1993, the ZZZZ authors were contributing significantly more to papers (as measured by their ordinal position in the authors list) than in 1997. The typical paper in the 1993 database had about three authors, with the ZZZZ author being second.
* Those papers with a VH figure of merit tended to have more ZZZZ lead authors than one would expect from the database overall average.
* While the more prolific ZZZZ authors in 1993 had better relative citeability than the database average, these authors were closer to the end of the author listing than the database average.
* More work needs to be done to place ordinal position quantification on a stronger scientific foundation.

In about half the cases, papers with a ZZZZ author address were cited as well as, or better than, comparable non-ZZZZ address papers. On the surface, it appears that papers with ZZZZ authors are having a reasonable impact on the technical community. However, the contribution of the ZZZZ authors to these papers, especially those where the ZZZZ author is listed last, remains unknown. It would have been useful to compare the number of authors for each paper in the pair; this might have shed some light on whether or not the ZZZZ papers are 'author heavy'. This was not done because this issue was not recognized until now. It would also be useful to ascertain why the ZZZZ authors dropped back in their ordinal position in the author list from 1993 to 1997.

VI. SUMMARY AND CONCLUSIONS

This appendix has provided two examples of the application of citation analysis to proposal evaluation. A number of lessons were learned concerning requirements for high quality citation analysis. These lessons are summarized as follows.

A. Since citation counts can vary sharply across sub-disciplines, absolute counts have little meaning, especially in the absence of absolute citation count performance standards. In order to provide meaning and context of citation counts for performance evaluation in citation analysis, some type of citation count normalization is required.

B. Three types of reference standards are used traditionally for citation analysis: 1) Reference standards based on prior sub-field classification; 2) Journals as reference standards; 3) Related records as reference standards.

None of the above normalization methods are adequate for precise normalization, since they do not provide sufficient resolution for distinguishing among the lower level sub-fields. Inability to distinguish precisely among sub-fields translates, in some cases, to substitution of far different magnitude numbers for the normalization base

C. Selection of papers in the SCI thematically similar to the target paper depends strongly on the study's purpose and objectives, the mission of the performing organization, the degree of focus of the paper's theme, the size of the research paper pool from which to choose, and the level of technical description in the paper's SCI Abstract. The relation to study purpose is especially important, and is often overlooked. If the focus is on 'job right' quality, then the thematically similar papers will be limited to a very narrow area of inquiry. If the focus is on 'right job' quality, then the focus of thematically related papers can be expanded greatly. The practical implications of 'job right' vs 'right job' comparisons are that papers with substantially higher citation counts could be included in the normalization pool as the allowed definition of thematic similarity becomes broadened.

D. Selection of papers thematically similar to the target paper was very difficult, time-consuming, and subjective. This was especially true for the broad-based analyses. The selection was more straightforward for the much more limited specific technology papers, since these more focused areas seemed to have many researchers working related problems. The subjectivity involved in selecting thematically similar papers is a major source of uncertainty of the results. A rigorous study, in addition to having the rigorous information retrieval and statistical sampling processes mentioned in the next two paragraphs, requires the use of multiple evaluators for the same target papers to average out bias.

E. Many of the applied research target papers combined analytical technique advancement with novel application advancement. It was not always possible to have thematic similarity for both technique and application, especially in those research areas with relatively few performers. Typically, a choice had to be made between technique and application for determining thematic similarity.

F. Two important issues were i) determining the number of thematically similar candidate papers in the pool from which to choose, and then ii) determining the number of papers to select from the pool. First, in a credible study, candidate thematically similar papers would be identified by the most rigorous processes available, and such processes are presently very complex and time-consuming. Second, the number of papers to select from the candidate pool should have the greatest thematic similarity, and be representative statistically. Such selection would have required poring over hundreds, or thousands, of similar papers, and selecting a substantial number of the most representative thematically.

G. Contrary to much popular thinking, the technical expertise of the citation analyst can have a major impact on the quality of the results. The type of pair-wise comparison required for credible citation studies is a highly subjective process, requiring the selection of a thematically similar normalization base. If the analyst understands the subject matter, the subjective judgements made will be reasonably accurate. If the analyst is not a technical expert in the subject area, the results will contain a high degree of uncertainty. Thus, in a rigorous citation analysis, multiple technical experts are necessary to average out individual bias and subjectivity, and much manually intensive effort is required for the normalization process.

Operationally, the above results suggest that a credible citation analysis for determining performer or team quality should have the following components:

*Multiple technical experts to average out individual bias and subjectivity
*A process for comparing performer or team output papers with a normalization base of similar papers
*A process for retrieving a substantial fraction of candidate normalization base papers
*Manual evaluation of many candidate normalization base papers to obtain high thematic similarity and statistical representation

Since the use of citation analysis as one metric for determining research performer or team quality is substantially under-utilized in government and industry at present, the addition of the above requirements to the citation analysis process would only serve to reduce its utilization further. Pragmatically, tradeoffs are required if citation analysis is to be used as an evaluative tool. The degradation in citation analysis quality as the above conditions are relaxed needs to be studied further.

VII. APPENDIX REFERENCES

Albert, M.B., Avery, D., Narin F., Mcallister P., “Direct Validation Of Citation Counts As Indicators Of Industrially Important Patents” , Research Policy 20: (3) 251-259 , Jun 1991.

Braam, R.R., Moed H.F., Vanraan A.F.J., “Mapping Of Science By Combined Co-Citation And Word Analysis .1. Structural Aspects” , Science Technology & Human Values 13: (1-2) 97-98 ,Win- Spr 1988.

Braam, R.R., Moed, H.F., Vanraan A.F.J.,“Mapping Of Science By Combined Cocitation And Word Analysis .1. Structural Aspects”, Journal Of The American Society For Information Science 42: (4) 233-251, May 1991.

Davidse, R. J., and VanRaan, A. F. J., “Out of Particles: Impact of CERN, DESY, and SLAC Research to Fields other than Physics”, Scientometrics, 40:2. P. 171-193, 1997.

Del Río, J. A., Kostoff, R. N., García, E. O., Ramírez, A. M., and Humenik, J. A., “Citation Mining Citing Population Profiling using Bibliometrics and Text Mining”. Centro de Investigación en Energía, Universidad Nacional Autonoma de Mexico, 2001. http://www.cie.unam.mx/W_Reportes.

DOE, "An Assessment of the Basic Energy Sciences Program", Office of Energy Research, Office of Program Analysis, Report No. DOE/ER-0123, March 1982.

Festschrift, ASIST Monograph Series, Web Of Knowledge - A Festschrift In Honor Of Eugene Garfield, 2000.

Garfield, E., “Citation Indexes For Science - New Dimension In Documentation Through Association Of Ideas”, Science, 122: (3159) 108-111, 1955.

Garfield , E., Sher , I. H., “New Factors In Evaluation Of Scientific Literature Through Citation Indexing” American Documentation, 14: (3) , 1963.

Garfield , E., “Science Citation Index-New Dimension In Indexing Unique Approach Underlies Versatile Bibliographic Systems For Communicating + Evaluating Information”, Science, 144: (361), 1964.

Garfield , E., “Can Citation Indexing Be Automated”, Statistical Association Methods For Mechanized Documentation Symposium Proceedings 1964: (Nbs26) 189, 1965.

Garfield E., “Patent Citation Indexing And Notions Of Novelty Similarity And Relevance”, Journal Of Chemical Documentation 6: (2), 1966.

Garfield., E., “Citation Indexing For Studying Science”, Nature 227: (5259) , 1970.

Geisler, E., “The Metrics of Science and Technology”, Quorum Books, Westport, CT, 2000.

Kostoff, R. N., "The Handbook of Research Impact Assessment," Seventh Edition, Summer 1997, DTIC Report Number ADA296021. Also, available at http://www.dtic.mil/dtic/kostoff/index.html, 1997a.

Kostoff, R. N., "Peer Review: The Appropriate GPRA Metric for Research", Science, Volume 277, 1 August 1997b.

Kostoff, R. N., "Citation Analysis Cross-Field Normalization: A New Paradigm", Scientometrics, 39:3, 1997c.

Kostoff, R. N., Eberhart, H. J., and Toothman, D. R., "Database Tomography for Information Retrieval", Journal of Information Science, 23:4, 1997d.

Kostoff, R. N., "The Use and Misuse of Citation Analysis in Research Evaluation", Scientometrics, 43:1, September, 1998.

Kostoff, R. N., “The Underpublishing of Science and Technology Results”, The Scientist, 1 May 2000a.

Kostoff, R. N., "High Quality Information Retrieval for Improving the Conduct and Management of Research and Development", Proceedings:

Twelfth International Symposium on Methodologies for Intelligent Systems, 11-14 October 2000b.

Kostoff, R. N., and Hartley, J., “Structured Abstracts For Technical Journals”, Science, 11 May 2001a.

Kostoff, R. N., and Hartley, J., “Structured Abstracts For Technical Journal Articles, Letter to Technical Journal Editors, 14 May 2000b. Letter available from author.

MacRoberts, M.H., and MacRoberts, B.R., "Problems of Citation Analysis: A Critical Review," Journal of the American Society for Information Science, 40:5, 1989.

MacRoberts, M., and MacRoberts, B., "Problems of Citation Analysis", Scientometrics, 36:3, July-August, 1996.

Marshako.Iv , “System Of Document Connections Based On References”, Nauchno-Tekhnicheskaya Informatsiya Seriya 2-Informatsionnye Protsessy I Sistemy, (6) 3-8 1973

Marshakova Iv , “Citation Networks In Information-Science”, Scientometrics, 3: (1) 13-25 1981

Marshakova Iv , “On The Mapping Of Science”, Vestnik Akademii Nauk Sssr, (5) 70-82 1988

Martin, B. and Irvine, J., "Assessing Basic Research: Some Partial Indicators of Scientific Progress in Radio-Astronomy," Research Policy, 12, 1983.

Moed H.F., Vanraan A.F.J., “Observations And Hypotheses On The Phenomenon Of Multiple Citation To A Research Groups Oeuvre.”, Scientometrics 10: (1-2) 17-33 , July 1986.

Narin. , F., Carpenter M.P., “National Publication And Citation Comparisons”, Journal Of The American Society For Information Science 26: (2) 80-93 , 1975.

Narin, F., 'Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity" (monograph), NSF C-637, National Science Foundation, Contract NSF C-627, NTIS Accession No. PB252339/AS, March 31, 1976.

Narin, F., Carpenter M.P., Woolf P., “Technological Performance Assessments Based On Patents And Patent Citations”, IEEE Transactions On Engineering Management 31: (4) 172-183 , 1984

Narin, F., Olivastro, D., and Stevens, K. A., "Bibliometrics -Theory, Practice, and Problems", in: Kostoff, R. N., (ed.), Evaluation Review, Special Issue on Research Impact Assessment, 18:1, February 1994.

Narin F., Hamilton K.S., “Bibliometric Performance Measures”, Scientometrics 36: (3) 293-310 , Jul-Aug 1996.

Nederhof A.J., Vanraan A,F.J., “Citation Theory And The Ortega Hypothesis”, Scientometrics 12: (5-6) 325-328 , Nov 1987

Schubert A., Glanzel W. , Braun T., “Relative Indicators Of Publication Output And Citation Impact Of European Physics Research: 1978-1980”, Czechoslovak Journal Of Physics 36: (1) 126-129 , 1986

Schubert A., Braun T., ”Reference-Standards For Citation Based Assessments”, Scientometrics 26: (1) 21-35 , Jan 1993

Schubert, A., and Braun, T., "Cross-Field Normalization of Scientometric Indicators", Scientometrics, 36:3, 1996.

Small , H. G. “Relationship Between Citation Indexing And Word Indexing - Study Of Co-Occurrences Of Title Words And Cited References”, Proceedings Of The American Society For Information Science 10: 217-218 , 1973.

Small , H., “Co-Citation In Scientific Literature - New Measure Of Relationship Between 2 Documents”, Current Contents (7) 7-10, 1974.

Small , H. G., “Co-Citation Model Of A Scientific Specialty - Longitudinal-Study Of Collagen Research”, Social Studies Of Science 7: (2) 139-166, 1977.

Small. , H., “The Relationship Of Information-Science To The Social-

Sciences - A Co-Citation Analysis", Information Processing & Management 17: (1) 39-50, 1981.

Small H., Sweeney E., Greenlee E., "Clustering The Science Citation Index Using Co-Citations .2. Mapping Science", Scientometrics , 8, 1985.

Sullivan D, White Dh, Barboni Ej, "Co-Citation Analyses Of Science – Evaluation", Social Studies Of Science, 7: (2) 223-240 1977

Sullivan D, Koester D, White Dh, Kern R, "Understanding Rapid Theoretical Change In Particle Physics - Month-By-Month Co-Citation Analysis," Proceedings Of The American Society For Information Science, 16: 276-285 1979

Sullivan D, Koester D, White Dh, Kern R, "Understanding Rapid Theoretical Change In Particle Physics - A Month-By-Month Co-Citation Analysis," Scientometrics, 2: (4) 309-319 1980

Vanraan, A.F.J., "Fractal Geometry of Information Space As Represented By Co-Citation-Clustering" , Scientometrics 20: (3) 439-449 , Mar-Apr 1991.

Vanraan, A.F.J., Tijssen R.J.W., "The Neural Net Of Neural Network Research - An Exercise In Bibliometric Mapping" , Scientometrics 26: (1) 169-192 , Jan 1993.

Vanraan, A.F.J., "Advanced Bibliometric Methods As Quantitative Core Of Peer Review Based Evaluation And Foresight Exercises", Scientometrics 36: (3) 397-420 , Jul-Aug, 1996.

Zachlin, A. C., "On Literature Citation", Science, 107: (2777) 292-293, 1948.

Zirkle, C., "Citation of Fraudulent Data", Science, 120: (3109) 189-190, 1954.

Zsindely, S., Schubert A., Braun T., "Citation Patterns Of Editorial Gatekeepers In International Chemistry Journals", Scientometrics 4: (1) 69-76 , 1982.