## ARTICLES

# Fullerene Data Mining Using Bibliometrics and Database Tomography

Ronald N. Kostoff*

Office of Naval Research, 800 North Quincy Street, Arlington, Virginia 22217

Tibor Braun

Institute of Inorganic and Analytical Chemistry, Eotvos Lorand University, Budapest, Hungary

Andras Schubert

Library of the Hungarian Academy of Sciences, Budapest, Hungary

Darrell Ray Toothman

RSIS, Inc., 1651 Old Meadow Road, Suite 500, Mclean, Virginia

James A. Humenik

Noesis, Inc., 4200 Wilson Boulevard, Suite 900, Arlington, Virginia 22203

Received May 10, 1999

Database tomography (DT) is a textual database analysis system consisting of two major components: (1) algorithms for extracting multiword phrase frequencies and phrase proximities (physical closeness of the multiword technical phrases) from any type of large textual database, to augment (2) interpretative capabilities of the expert human analyst. DT was used to derive technical intelligence from a fullerenes database derived from the Science Citation Index and the Engineering Compendex. Phrase frequency analysis by the technical domain experts provided the pervasive technical themes of the fullerenes database, and phrase proximity analysis provided the relationships among the pervasive technical themes. Bibliometric analysis of the fullerenes literature supplemented the DT results with author/journal/institution publication and citation data. Comparisons of fullerenes results with past analyses of similarly structured near-earth space, chemistry, hypersonic/supersonic flow, aircraft, and ship hydrodynamics databases are made. One important finding is that many of the normalized bibliometric distribution functions are extremely consistent across these diverse technical domains and could reasonably be expected to apply to broader chemical topics than fullerenes that span multiple structural classes. Finally, lessons learned about integrating the technical domain experts with the data mining tools are presented.

## 1. INTRODUCTION

Science and technology are assuming an increasingly important role in the conduct and structure of domestic and foreign business and government. In the highly competitive civilian and military worlds, there has been a commensurate increase in the need for scientific and technical intelligence to ensure that one's perceived adversaries do not gain an overwhelming advantage in the use of science and technology. While there is no substitute for direct human intelligence gathering, there have become available many techniques that can support and complement it. In particular, techniques that identify, select, gather, cull, and interpret large amounts of technological information semiautomatically can expand greatly the capabilities of human beings in performing technical intelligence.

One such technique is database tomography (DT),[1−3] a system for analyzing large amounts of textual computerized

* To whom correspondence should be addressed. Phone: 703-696-4198. FAX: 703-696-4274. E-mail: KOSTOFR@ONR.NAVY.MIL.

10.1021/ci990045n CCC: $19.00 
Published on Web 12/08/1999

material. It includes algorithms for extracting multiword phrase frequencies and phrase proximities from the textual databases, coupled with the topical expert human analyst to interpret the results and convert large volumes of disorganized data to ordered information. Phrase frequency analysis (occurrence frequency of multiword technical phrases) provides the pervasive technical themes of a database, and phrase proximity (physical closeness of the multiword technical phrases) analysis provides the relationships among pervasive technical themes, as well as among technical themes and authors/journals/institutions/countries, etc. The present paper describes use of the DT process, supplemented by literature bibliometric analyses, to derive technical intelligence from the published literature of fullerene science and technology.

Fullerene, as defined by the authors for this study, consists of theory/experiment/computation/applications related to large ordered carbon atom clusters. It is defined operationally by the following query, obtained by the iterative technique referenced in the next paragraph:

"C-60" OR "C-70" OR "C60" OR "C70" OR
FULLERENE* OR CARBON NANOTUBE* OR
BUCKMINSTERFULLERENE OR
FULLERIDE* OR FULLERITE OR
METALLOFULLERENE* OR
METHANOFULLERENE OR ENDOHEDRAL OR
SOCCERBALL OR BUCKEYTUBE* OR "C-78"

To execute the study reported in this paper, a database of relevant fullerene articles is generated using the iterative search approach of simulated nucleation.[4,5] Then, the database is analyzed to produce the following characteristics and key features of the fullerene field: recent prolific fullerene authors; journals that contain numerous fullerene papers; institutions that produce numerous fullerene papers; keywords most frequently specified by the fullerene authors; authors whose works are cited most frequently; particular papers and journals cited most frequently; pervasive themes of fullerene; and relationships among the pervasive themes and subthemes. Finally, the lessons learned from this study (and two parallel studies) from integrating the topical domain experts with the analytical data mining tools are summarized.

What is the importance of applying DT and bibliometrics to a topical field such as fullerenes? The road map, or guide, of this field produced by DT and bibliometrics provides the demographics and a macroscopic view of the total field in the global context of allied fields. This allows specific starting points to be chosen rationally for more detailed investigations into a specific topic of interest. DT and bibliometrics do not obviate the need for detailed investigation of the literature or interactions with the main performers of a given topical area in order to make a substantial contribution to the understanding or the advancement of this topical area, but they allow these detailed efforts to be executed more efficiently. DT and bibliometrics are quantity-based measures (number of papers published, frequency of technical phrases, etc.), and correlations with intrinsic quality are less direct. The direct quality components of detailed literature investigation and interaction with performers, combined with the DT and bibliometrics analysis, can result in a product highly relevant to the user community.

## 2. BACKGROUND

**2.1. Overview.** The information sciences background for the approach used in this paper is presented in Kostoff.[6] This reference shows the unique features of the computer and co-word-based DT process relative to other road map techniques. It describes the two main road map categories (expert-based and computer-based), summarizes the different approaches to computer-based road maps (citation and cooccurrence techniques), presents the key features of classical co-word analysis, and shows the evolution of DT from its co-word roots to its present form.

**2.2. Development of DT.** Classical co-word analysis applied to index/keywords for the purpose of science and technology (S&T) evaluation does not allow the richness of the semantic relationships in full text to be exploited, and it is restricted to formally published papers. To allow any form of free text to be used, DT was developed.

In 1990−1991, experiments were performed at the Office of Naval Research that showed the frequency with which phrases appeared in full text narrative technical documents was related to the main themes of the text. The phrases with the highest frequencies of appearance represented the main, "pervasive" themes of the text. In addition, the experiments showed that the physical proximity of the phrases was related to the thematic proximity. These experiments formed the basis of DT.

The DT method in its entirety requires generically three distinct steps. The first step is identification of the main themes of the text being analyzed. The second step is determination of the quantitative and qualitative relationships among the main themes and their secondary themes. The final step is tracking the evolution of these themes and their relationships through time. The first two steps will be summarized now. Time evolution of themes has not yet been studied.

First, the frequencies of appearance in the total text of all single word phrases (e.g., matrix), adjacent double word phrases (e.g., metal matrix), and adjacent triple word phrases (e.g., metal matrix composites) are computed. The highest frequency significant technical content phrases are selected by topical experts as the pervasive themes of the full database.

Second, for each theme phrase, the frequencies of phrases within $\pm M$ (nominally 50) words of the theme phrase are computed for every occurrence of the theme phrase in the full text, and a phrase frequency dictionary is constructed. This dictionary contains the phrases closely related to the theme phrase. Numerical indices are employed to quantify the strength of this relationship. Both quantitative and qualitative analyses are performed by the topical expert for each dictionary (hereafter called cluster) yielding, among many results, those subthemes closely related to and supportive of the main cluster theme.

Third, threshold values are assigned to the numerical indices, and these indices are used to filter out the phrases most closely related to the cluster theme. However, because numbers are limited in their ability to portray the conceptual relationships among themes and subthemes, the qualitative

analyses of the extracted data by the topical experts have been at least as important as the quantitative analyses. The richness and detail of the extracted data in the full text analysis allow an understanding of the theme interrelationships not heretofore possible with previous text abstraction techniques (using index words, keywords, etc.).

At this point, a variety of different analyses can be performed. For databases of nonjournal technical articles,[1] the final results have been identification of the pervasive technical themes of the database, the relationship among these themes, and the relationship of supporting subthrust areas (both high and low frequency) to the high-frequency themes. For the more recent studies in which the databases are journal article abstracts and associated bibliometric information (authors, journals, addresses, etc.), the final results have also included relationships among the technical themes and authors, journals, institutions, etc.[7]

The study reported in the present paper is in the latter (journal article abstract) category. It differs from the previous published papers in this category[6−8] in four respects. First, the topical domain (fullerenes) is completely different. Second, a much more comprehensive bibliometrics cross-discipline comparison is performed. Third, the balance of effort has shifted from computer-centric (where the primary emphasis was on the computer results, and the secondary emphasis was on the expert analysis of the computer results) to expert-centric (where the primary emphasis is on expert analysis of the computer results and raw data, and the computer results serve to augment the capabilities of the expert). There are two reasons for this shift in emphasis. Expert-centric S&T data mining provides an in-depth understanding/identification of the technical concepts and their interrelationships, whereas the computer-centric approach focused on the more superficial level of context-free phrases. Also, as shown in later sections of this paper, one of the major products of a serious data mining study is the "educated expert", who has had his/her horizons broadened substantially by the data mining experience. The study experience should center around maximum enhancement of the capabilities of the expert in the topical area. Fourth, the study describes the data mining lessons learned from focusing on the integration of the technical domain expert with the computational tools.

**2.3. Evolution of DT into Textual Data Mining.** Recent evaluations of real-world textual data mining applications (unpublished) across a number of organizations showed a strong decoupling of the research performer from the data mining user. The performer tended to focus on the development of exotic automated techniques, to the relative exclusion of the components of judgment necessary for user credibility and acceptance. Consequently, the data mining techniques actually employed by most of the potential users examined were reading of copious numbers of articles obtained by the simplest of queries.

The DT process reported in this paper represents the framework of a data mining approach that will couple the data mining research and associated computer technology processes much more closely with the data mining user. Strategic database maps will be developed on the front end of the process using bibliometrics and DT, with heavy involvement from topical domain experts (either users or their proxies) in the DT component of strategic map generation. The strategic maps themselves will then be used as guidelines for detailed expert analysis of segments of the total database. The authors believe that this is the proper use of automated techniques for data mining: to augment and amplify the capabilities of the expert by providing insights to the database structure and contents, not to replace the experts by a combination of machines and nonexperts.

## 3. DATABASE GENERATION

The key step in the fullerene literature analysis is the generation of the database. For the present study, two databases were used.

**3.1. Science Citation Index.[9]** The first database consists of selected journal records (including authors, titles, journals, author addresses, author keywords, abstract narratives, and references cited for each paper) obtained by searching the web version of the Science Citation Index (SCI) for fullerene articles. At the time the present paper was written (late 1998), the version of the SCI used accessed about 5300 journals (mainly in physical, engineering, and life sciences basic research).

The SCI database selected represents a fraction of the available fullerene (mainly research) literature. It does not include the large body of classified literature or company proprietary technology literature. It does not include technical reports or books or patents on fullerenes. It covers a finite slice of time (1991 to mid-1998). The database used represents the bulk of the peer-reviewed high-quality fullerene science and technology and is a representative sample of all fullerene science and technology in recent times.

To extract the relevant articles from the SCI, the title, keyword, and abstract fields were searched using keywords relevant to fullerenes, although different procedures were used to search the title and abstract fields.[4] The resultant abstracts were culled to those relevant to fullerenes. The search was performed with the aid of two powerful DT tools (multiword phrase frequency analysis and phrase proximity analysis) using the process of simulated nucleation.[4]

An initial query of FULLERENE* and related terms produced two groups of papers: one group was judged by domain experts to be relevant to the subject matter; the other was judged to be nonrelevant. Gradations of relevancy or nonrelevancy were not considered. An initial database of titles, keywords, and abstracts was created for each of the two groups of papers. Phrase frequency and proximity analyses were performed on this textual database for each group. The high-frequency single, double, and triple word phrases characteristic of the relevant group, and their boolean combinations, were then added to the query to expand the papers retrieved. Similar phrases characteristic of the nonrelevant group were effectively subtracted from the query to contract the papers retrieved. The process was repeated on the new database of titles, keywords, and abstracts obtained from the search. A few more iterations were performed until the number of records retrieved stabilized (convergence).

The final query used for the fullerene study, shown in the Introduction, contained 15 terms. In other studies, such as aircraft S&T, the final query contained over 200 terms. There are two main reasons for the difference in query complexity. First, in the aircraft study, the coverage is much broader than

in the fullerene study. Second, but perhaps more importantly, the contents of the SCI database are more aligned with the objectives of the fullerene study than those of the aircraft study. As will be shown later by the results, the journal literature on fullerenes describes a research field well-aligned with the contents of the SCI research database. Aircraft is both a science/technology area as well as a tool/platform for performing research. While the SCI is well-aligned with the science/technology component of aircraft (e.g., aircraft structures, aircraft propulsion), the SCI also includes papers relating to the use of aircraft as a platform from which to perform research (e.g., crop spraying, buffalo tracking). If the search philosophy is to start the iterative query process with AIRCRAFT and subtract terms not applicable to the platform function of aircraft, then a large SCI query will be required for aircraft to remove these platform-oriented terms. This type of dual usage does not exist yet for fullerenes in the published journal literature and is therefore reflected in the much simpler fullerene query.

The situation is analogous to selection of a mathematical coordinate system for solving a physical problem. If the coordinate system is aligned naturally with the body geometry (e.g., a spherical coordinate system used to model flow around a sphere), then a minimal number of equation terms is necessary. If the coordinate system is mismatched to the body geometry (e.g., a spherical coordinate system used to model the flow around a parallelepiped), then a large number of equation terms will be required to effectively translate between the two geometries.

The authors believe that queries of these magnitudes and complexities are required when necessary to provide a tailored database of relevant records that encompasses the broader aspects of target disciplines. In particular, if it is desired to enhance the transfer of ideas across disparate disciplines, and thereby stimulate the potential for innovation and discovery from complementary literatures,[10] then even more complex queries using simulated nucleation may be required.

The authors believe that the "purity" and completeness of the database of topically relevant records obtained using simulated nucleation is a key reason that the invariance of most of the normalized bibliometric distributions across different topical domains can be displayed (see the normalized bibliometric distribution functions in later sections). One beneficial value of utilizing simulated nucleation is that the search terms are obtained from the words of the authors in the SCI and Engineering Compendex (EC) databases, not by guessing on the part of the searcher.

**3.2. Engineering Compendex.[11]** The second database consists of selected journal and conference proceeding records (including authors, titles, journals, author addresses, author keywords, abstract narratives, and references cited for each paper) obtained by searching the CD-ROM version of the Engineering Compendex for fullerene articles. In late 1998, this version of the EC accessed about 2600 journals, mainly in physical and engineering sciences applied research and technology).

The EC database selected represents a fraction of the available fullerene (mainly applied research and technology) literature. It does not include either the large body of classified and company proprietary technology literature or the large body of technical reports on fullerenes. It covers a finite slice of time (1991 to mid-1998). Because of the monolithic research nature of fullerenes, the same query used for searching the SCI was used to search the EC.

## 4. RESULTS

The results from the publications bibliometric analyses are presented in section 4.1, followed by the results from the citations bibliometrics analysis in section 4.2. Results from the DT analyses are shown in section 4.3. The SCI and EC bibliometric fields incorporated into the database included, for each paper, the author, journal, institution, and keywords. In addition, the SCI included references for each paper. Due to the fundamental research orientation of fullerenes as reflected in the published journal literature used for this study, most of the EC results were included in the SCI results. Therefore, only the SCI results will be presented in this paper.

The bibliometrics sections (4.1, 4.2) have two components. Important numerical indicators are presented that illuminate some aspect of the fullerenes research literature (e.g., average authors per paper, number of journals, papers per institution), and distribution functions of publication and citation parameters (e.g., numbers of authors $f(n)$ who publish "$n$" papers) are compared with those of other technical discipline studies that used a similar approach.

The DT sections contain three components. First, the high-frequency keywords are grouped into "natural" categories, and the picture they provide of the fullerenes literature (research, open literature, unclassified, nonproprietary) is described. Second, the high-frequency phrases from the abstracts are grouped into "natural" categories, and the picture they provide of the fullerenes literature is presented. Third, the high numerical indicator phrases from the proximity analyses of the abstracts and other portions of the database (author names, article titles, journal names, author addresses) are grouped into "natural" categories, and the picture they provide of the fullerenes literature is shown.

The meaning of the term "natural" is that these categories were not prescribed beforehand. From observation of the hundreds of different phrases and their frequencies, categories useful for interpreting and describing the main literature findings appeared to emerge.

The analytical approaches taken for the first three components (keyword phrase frequency, abstract phrase frequency, phrase proximity) are based on their fundamental data structures. The keyword and abstract phrase frequencies are essentially quantity measures. They lend themselves to "binning" and addressing adequacies and deficiencies in levels of effort. They do not contain relational information and therefore offer little insight into S&T linkages.

The phrase proximity results are essentially relational measures, although some of the proximity results imply levels of effort that support specific S&T areas. The phrase proximity results mainly offer insight into S&T linkages and have the potential to help identify innovative concepts from disparate disciplines.[10] Thus, the keyword and abstract phrase frequency analyses will be addressed to adequacy of effort, and the phrase proximity analyses will be addressed to relationships primarily and supporting levels of effort secondarily.

**4.1. Publication Statistics on Authors, Journals, Organizations, and Countries.** The first group of metrics presented is counts of papers published by different entities. These metrics can be viewed as output and productivity measures. They are not direct measures of research quality,

**Table 1.** DT Studies of Topical Fields

| metric/study | FUL | JACS | NES | HYD | HSF | AIR | RIA |
|---|---|---|---|---|---|---|---|
| no. of articles | 10 515 | 2150 | 5481 | 4608 | 1284 | 4346 | 2300 |
| start year | 1991 | 1994 | 1993 | 1991 | 1993 | 1991 | 1991 |
| end year | M-1998 | 1994 | M-1996 | M-1998 | M-1996 | M-1998 | E-1995 |

**Table 2.** Author Bibliometrics−SCI

| metric/study | FUL | JACS | NES | HYD | HSF | AIR | RIA |
|---|---|---|---|---|---|---|---|
| no. of authors | 12 837 | 6535 | 12 453 | 7 869 | 2483 | 6619 | 2975 |
| no. of author listings | 41 167 | 8151 | 18 474 | 10 558 | 3372 | 9085 | 3868 |
| av no. of listings per author | 3.2 | 1.2 | 1.5 | 1.3 | 1.38 | 1.4 | 1.3 |
| no. of papers retrieved | 10 515 | 2150 | 5481 | 4608 | 1284 | 4346 | 2300 |
| av no. of author listings per paper | 3.92 | 3.79 | 3.37 | 2.29 | 2.63 | 2.09 | 1.68 |

although there is some threshold quality level inferred due to these papers' publication in the (typically) high caliber of journals accessed by the SCI.

**4.1.1. Prolific Authors.** The author field was separated from the database, and a frequency count of author appearances was made. In the SCI database results, there were 12 839 different authors and 41 167 author listings (the occurrence of each author's name on a paper is defined as an author listing). While the average number of listings per author is about 3.2, the most prolific authors (e.g., ACHIBA Y, 143; KROTO HW, 121; KIKUCHI K, 115; SAITO Y, 112; TAYLOR R, 111; SHINOHARA H, 107; SMALLEY RE, 98) have listings about an order of magnitude greater than the average. There were 10 515 papers retrieved, yielding an average of 3.92 authors per paper.

Previous DT/bibliometrics studies were conducted of the technical fields of (1) near-earth space (NES)[7] (2) hypersonic and supersonic flow over aerodynamic bodies (HSF),[6] (3) chemistry (JACS)[8] as represented by *The Journal of the American Chemical Society*, (4) aircraft (AIR), and (5) hydrodynamic flow over surfaces (HYD). Overall parameters of these studies are shown in Table 1.

These studies yielded the following: (1) 3.37 authors per paper for the NES results; (2) 2.63 authors per paper for the HSF results; (3) 3.79 authors per paper for the chemistry results; (4) 2.09 authors per paper for the AIR results; (5) 2.29 authors per paper for the HYDRO results. A previous study on the nontechnical field of research impact assessment (RIA) yielded about 1.68 authors per paper. See Table 2 for summary statistics of these previous studies.

Table 2 compares the SCI author bibliometric statistics for the different studies. These studies are listed, proceeding from left to right, in approximate order of the (subjectively estimated) science/technology ratio of the underlying field. Thus, the leftmost field listed, FUL, is estimated to be the most basic (based on the specific query used and the themes of the papers retrieved), and the rightmost technical field, AIR, is estimated as the most applied. RIA, the rightmost column, is not a technical field and is listed for completeness only. It should be emphasized that the subjective judgments used to estimate the maturity of these technical fields were based on the SCI journal papers only and not on other data sources such as patent databases.

In Table 2, five variables/figures of merit are presented for each study. The number of authors represents the total number of different names contained in the author blocks, while the number of author listings is the sum over all authors of the number of times each author's name was listed in an author block. The average number of (author) listings per author is the ratio of the above two quantities. The number of papers retrieved is the total number of relevant papers that comprised the database and was used for the analyses, while the average number of author listings per paper is the number of author listings divided by the number of papers retrieved.

In all cases, the most prolific authors had listings more than an order of magnitude greater than the average number of listings per author. The average number of listings per author is remarkably consistent except for FUL, where it is about 2.5 times the average of the other fields studied. FUL is a very young and dynamic research field, with extensive global activity, participation, and competition. Based on the SCI and EC papers examined for the present study, there is little technology development at present, at least in comparison with the other fields. Whereas the technology component of myriad fields tends to be characterized by less papers than the research component, FUL does not suffer from this limitation on its average activity. In addition, for developed S&T areas, many of the papers may not have a strict discipline focus but may address uses of the technology. These papers could be somewhat peripheral or tangential to the central discipline, and the authors may not be heavy contributors to the discipline per se. In FUL, the papers are written by active researchers solely focused on advancing the state-of-the-art, and the peripheral authors who might contribute a paper ot two do not surface often in this topical research area.

While there is a wide range among disciplines in the number of papers retrieved, the average number of author listings per paper decreases steadily, proceeding from the most basic fields to the most applied. The three most basic fields (FUL, JACS, NES) tend to be experiment-dominated, with much less effort devoted to computational modeling (as will be shown in the later DT sections). In many cases, these experiments require expensive equipment and large teams of researchers because of their complexity, and this is reflected in the large numbers of authors on the papers produced.

Figure 1 shows the distribution function of author listing frequency for the fullerene, NES, JACS, HSF, AIR, and HYDRO databases. The abcissa is the number of author listings $n$, and the ordinate is the number of authors $f(n)$ who have author listing $n$. In each case, the distribution function has been normalized to the number of authors who have one listing in the respective databases. The graph is plotted on a semilog scale to stretch the lower ordinate region.

The solid line in Figure 1 is the nominal "$1/n^2$" Lotka's law[12] distribution. With the exception of the FUL data, all

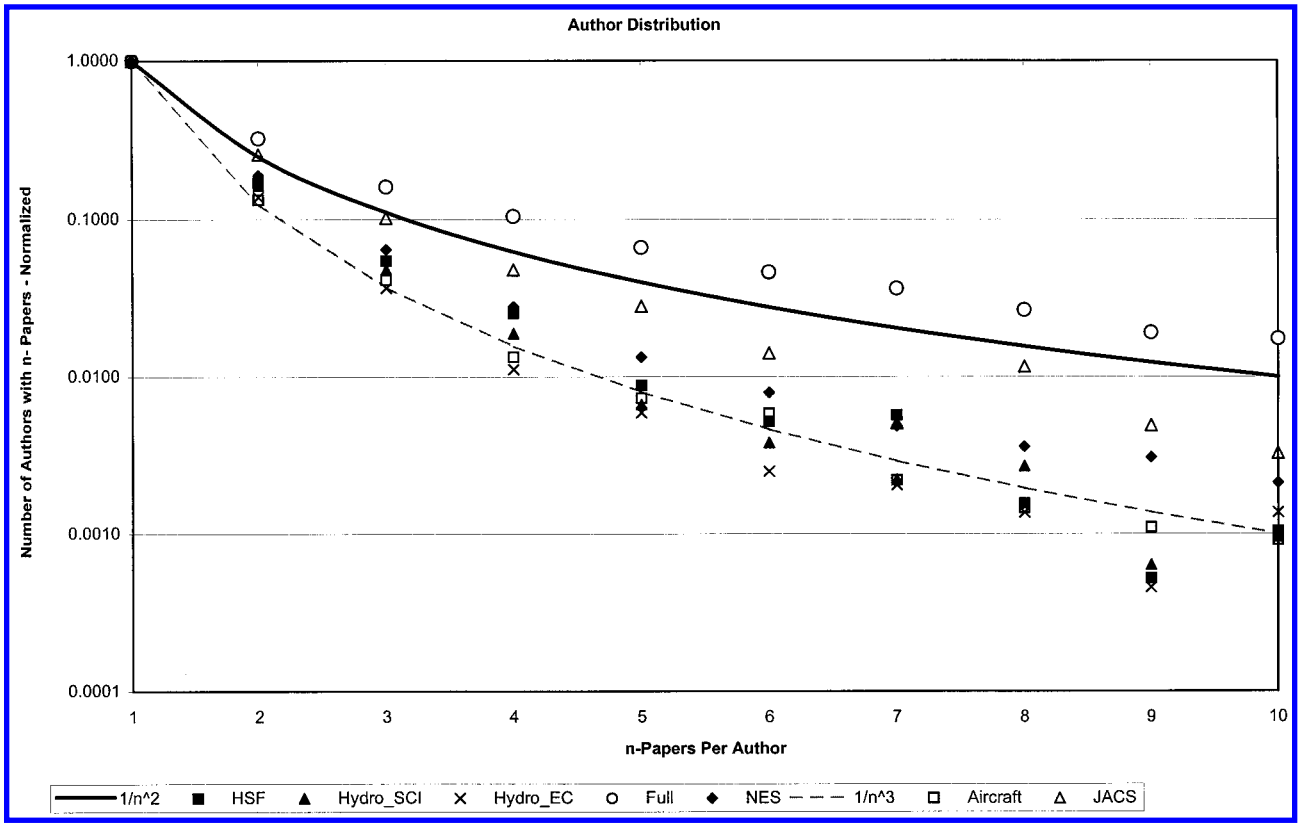


**Figure 1.**

of the experimental data decline much steeper than the $1/n^2$ law predicts, centering about a $1/n^3$ distribution. In the studies reported in the present document, the base of journals has been widened relative to what was available to Lotka. More journals of all types are available through the SCI. Also, because of the S&T scope of the present studies, more technology, and applications-oriented journals of peripheral relation to the core science disciplines are included. As the base of journals is widened, and more noncore journals are included in the source database, a larger diversity of authors is also included in the source database. These additional authors, who are less prolific and recognized in the discipline than the core authors, will populate the lower regions of the distribution function and will effectively skew the distribution function toward larger gradients relative to the Lotka distribution.

In the anomalous FUL case, the discipline is sufficiently young and mainly in the basic research phase that the widening of the journal base has not yet occurred. As the next section on journal bibliometrics shows, even though FUL has twice the numbers of papers relative to any of the other fields examined in this study, the total number of journals in which FUL authors publish is no larger than any of the other fields. The research authors want to establish their reputations in the core research journals and therefore have a higher number of papers per journal as also shown in the next section. In addition, the more sporadic nature of publication in the discipline-peripheral technology and applications oriented journals has not yet occurred. The FUL case matches most closely the discipline structure used in Lotka's work, and the FUL distribution matches the nominal Lotka law distribution most closely.

In summary, the nominal Lotka distribution can be viewed as most applicable to core discipline authors associated with the core discipline literature, while the present method reported in this paper is more focused on studying the technical discipline from a broader perspective. In this sense, the specific form of Lotka's Law that applies then becomes a function of how one defines the literature and core journals in a field, as well as the development status of the discipline.

**4.1.2. Journals Containing Most Fullerene Papers.** A similar process was used to develop a frequency count of journal appearances. In the SCI database, there were 680 different journals represented, with an average of 15.5 papers per journal. The journals containing the most fullerene papers (e.g., CHEMICAL PHYSICS LETTERS, 800; PHYSICAL REVIEW B—CONDENSED MATTER, 780; JOURNAL OF PHYSICAL CHEMISTRY, 390; SYNTHETIC METALS, 341; FULLERENE SCIENCE AND TECHNOLOGY, 332; JOURNAL OF THE AMERICAN CHEMICAL SOCIETY, 302) had in some cases an order of magnitude more papers than the average.

Table 3 compares the SCI journal bibliometric statistics for the different studies. Four variables/figures of merit are presented for each study. The number of journals represents the total number of different journal names contained in the source blocks. The average number of papers per journal is the ratio of total papers retrieved to total number of journals. The Bradford's law[13] metric derives from the following definition/restatement of the law: if the journals for a bibliography are grouped in order of decreasing publications, such that each group of journals contains the same number of papers, then the ratio of the number of journals in each successive group will be a constant greater than unity. The Bradford's law metric in Table 3 is this ratio between journal groups.

In all of the studies performed, the journals containing the most papers had an order of magnitude more papers than

**Table 3.** Journal Bibliometrics—SCI

| metric/study | FUL | JACS | NES | HYD | HSF | AIR | RIA |
|---|---|---|---|---|---|---|---|
| no. of papers retrieved | 10 515 | 2150 | 5481 | 4608 | 1284 | 4346 | 2300 |
| no. of journals | 680 | 1 | 628 | 675 | 277 | 713 | 645 |
| av no. of papers per journal | 15.46 | 2150 | 8.73 | 6.83 | 4.6 | 6.10 | 3.57 |
| Bradford's law—ratio between groups | 2.2 | | 2 | 1.5 | 3 | 3.1 | |

the average number of papers per journal. One unexpected finding is the closeness of the magnitudes of the number of journals for the different studies. Of the seven different topics studied, using different experts and different queries and different versions of the SCI and having different science/technology ratios, the total number of journals for five of those topics is within about 10% of 650. In fact, for four of those five journals, the total number of journals is within about 5% of 650. There are two outliers, JACS and HSF. The JACS study used one year's issues from *The Journal of the American Chemical Society*, and HSF is a much narrower and more limited field than the other broader fields studied. The question arises, why would the total number of journals across diverse fields be so similar, especially since the total number of papers differed by about a factor of 5 for the five fields of interest? No obvious answer emerges.

The average number of papers per journal decreases as the topical areas become more applied. This reflects the reality that technology-oriented papers tend to be published in a greater variety of journals that have a smaller concentration about any single research discipline, whereas research-oriented papers tend to be published in a smaller group of journals that are heavily discipline focused. Before discussing the Bradford's law results for Table 3, examples of how the Bradford's law ratios are computed for HSF and FUL are presented below.

For the HSF database, the first journal group selected contained one journal with 231 papers (AIAA JOURNAL); the second group had 3 journals with 237 papers; the third group 9 journals with 229 papers; the fourth group 25 journals with 229 papers; and the fifth group 70 journals with 229 papers. The ratio of numbers of journals per group between successive groups was approximately 3, in excellent agreement with Bradford's law.

For the FUL database, the first group selected contained two journals with 1580 papers (CHEMICAL PHYSICS LETTERS, 800; PHYSICAL REVIEW B—CONDENSED MATTER, 780); the second group had 5 journals with 1627 papers; the third group 10 journals with 1642 papers; the fourth group 21 journals with 1584 papers; and the fifth group 47 journals with 1572 papers. The ratio of the numbers of journals per group between successive groups is approximately 2.2, again in agreement with Bradford's law.

For the Bradford's law results of Table 3, the basic fields tend to have a ratio of about 2, while the more applied fields have a ratio of about 3. This means that in the basic fields there are more core discipline-oriented journals in which researchers would be motivated to publish relative to those in the applied fields. This conclusion is substantiated further by a more detailed examination of the numbers presented in the FUL and HSF examples. For the first three journal groups, the ratio of the cumulative number of journals to the total number of journals for the topical area is 0.025 for FUL and 0.047 for HSF. Since the first two or three journal groups tend to be the core topical groups, this result means that there is more depth in the FUL core than in the HSF core. The journals in which researchers are motivated to publish penetrates much deeper into the total FUL journal body relative to the total HSF body. In other words, there are more good basic research journals available for publication in FUL than there are in HSF.

Figure 2 shows the distribution function of journal frequency for the fullerene, AIR, HYDRO, HSF, NES, and RIA databases. The JACS database was derived from one journal only, *The Journal of the American Chemical Society*, and therefore was not applicable to this chart. The abcissa is the number of papers $n$ from the relevant database published in a given journal, and the ordinate is the number of journals which contain $n$ papers. In each case, the distribution function has been normalized to the number of journals that contain one relevant paper. Again, because of the strong initial gradients, the graph is plotted on a semilog scale.

The solid line in Figure 2 is a $1/n^2$ distribution, and represents a lower bound of all the experimental data. On average, the FUL data again appear to have the shallowest gradients. The rationale follows that of the previous section and need not be repeated here.

**4.1.3. Institutions Producing Most Fullerene Papers.** A similar process was used to develop a frequency count of institutional address appearances. It should be noted that many different organizational components may be included under the single organizational heading (e.g., Harvard Univ could include the Chemistry Department, Biology Department, Physics Department, etc.). Lack of space precluded printing out the components under the organizational heading.

There were 2168 different organizations listed in the SCI author address organizations, with an average of 4.85 papers per organization. The institutions producing most fullerene papers (e.g., RUSSIA,RUSSIAN ACAD SCI, 602; USA, RICE UNIV, 467; USA, UNIV PENN, 314; USA, UNIV CALIF SANTA BARBARA, 264; UK, UNIV SUSSEX, 248; USA, MIT, 221; JAPAN, TOKYO METROPOLITAN UNIV, 217; JAPAN, TOHOKU UNIV, 207; PEOPLES R CHINA, CHINESE ACAD SCI, 206) were greater than an order of magnitude more productive than the average. In aggregate, the University of California campuses are the most productive of any of the institutions in terms of papers published (~700), although no statements can be made about their production efficiency, since research expenditures were not included in this study. The top position of the Russian Academy of Sciences and the high ranking of some Japanese universities and that of the Chinese Academy has to be considered remarkable.

Table 4 compares the SCI institutional bibliometric statistics for the different studies. Four variables/figures of merit are presented for each study. The number of institutions represents the total number of different institution names contained in the address blocks. The average number of papers per institution is the ratio of total papers retrieved to total number of institutions. The average number of authors

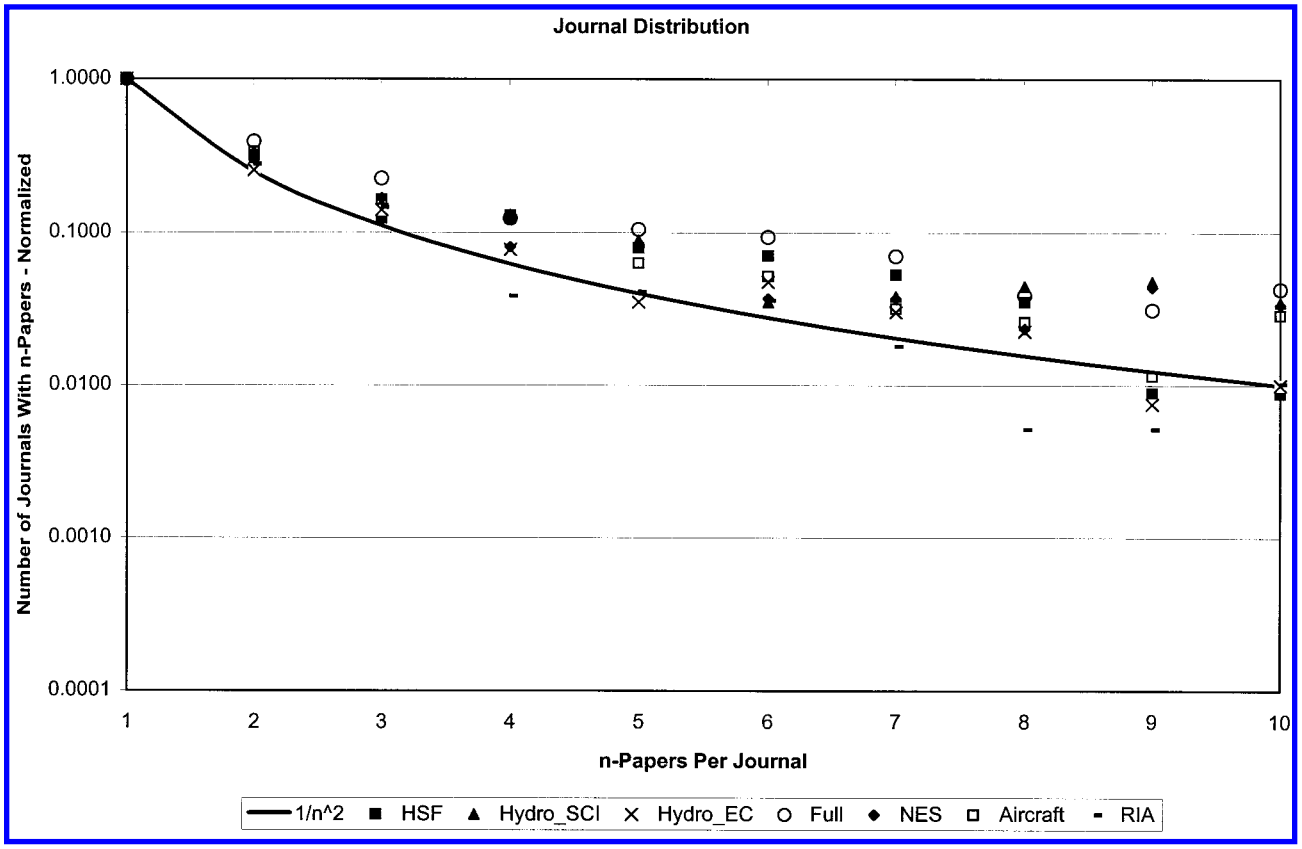


**Figure 2.**

**Table 4.**

| metric/study | FUL | JACS | NES | HYD | HSF | AIR | RIA |
|---|---|---|---|---|---|---|---|
| | | | (a) Institution Bibliometrics—SCI | | | | |
| no. of papers retrieved | 10 515 | 2150 | 5481 | 4608 | 1284 | 4346 | 2300 |
| no. of institutions | 2168 | 750 | 10 435 | 1905 | 661 | 1484 | 1125 |
| av no. of papers per institution | 4.85 | 2.9 | 0.53 | 2.42 | 1.94 | 2.93 | 2 |
| no. of authors per institution | 5.92 | 8.7 | 1.19 | 4.13 | 3.76 | 4.46 | 2.64 |
| | | | (b) Country Bibliometrics—SCI | | | | |
| no. of countries | 63 | 44 | 105 | 78 | 53 | 56 | 56 |
| ratio of US papers to five nearest producers | 0.73 | 2.5 | 1.94 | 1.32 | 1.6 | 1.74 | 2.47 |

per institution is the ratio of the total number of authors to the total number of institutions.

In all topical areas examined, the institutions producing the most papers were greater than an order of magnitude more productive than the average institution. The total number of institutions producing papers differs substantially for the different topical areas, with the NES number of institutions appearing as a major outlier. The average number of papers per institution does not follow any discernible trend, at least with respect to the science/technology ratio of the discipline. The NES average papers number is much lower than for the other topical areas. Combining the average author listings per paper result from Table 2 with the average papers per institution from Table 4, the NES picture is one of many diverse participants per study from myriad institutions.

For the near-earth space focus of the NES study, which centered mainly about unmanned satellites and the manned orbiting platforms, the space vehicle tends to serve as a "truck" or "bus", which transports the science experiments and scientists. Thus, the central NES component is not so much a technical research discipline as it is the vehicle that enables the research to be accomplished. The actual research performed is not focused on the vehicle and is spread among many very diverse areas and performers and institutions.

At the other extreme in Table 4, the number of papers per institution for FUL appears to be substantially greater than for the other studies. The dominant cause appears to derive from the large number of papers per author for FUL shown in Table 2. FUL is a young dynamic field with a number of centers containing strong efforts in this topical area (see last metric in Table 4), and the combination of high critical mass fractions per center with high productivity per author produces the large number of papers per institution.

There appear to be no discernible trends in Table 4 for the final metric, average number of authors per institution. Again, the NES value of 1.19 is substantially lower than that of the other studies, for the same reason that the number of papers per institution was lower. And again, using the NES EC results[7] of 14 036 authors and 2000−2700 institutions, the EC average of ~6.5 authors per institution is much more in line with the results of other studies in Table 4.

Figure 3 shows the distribution function of institution frequency for the fullerene, HSF, NES, JACS, AIR, and HYDRO databases. The abcissa is the number of papers n

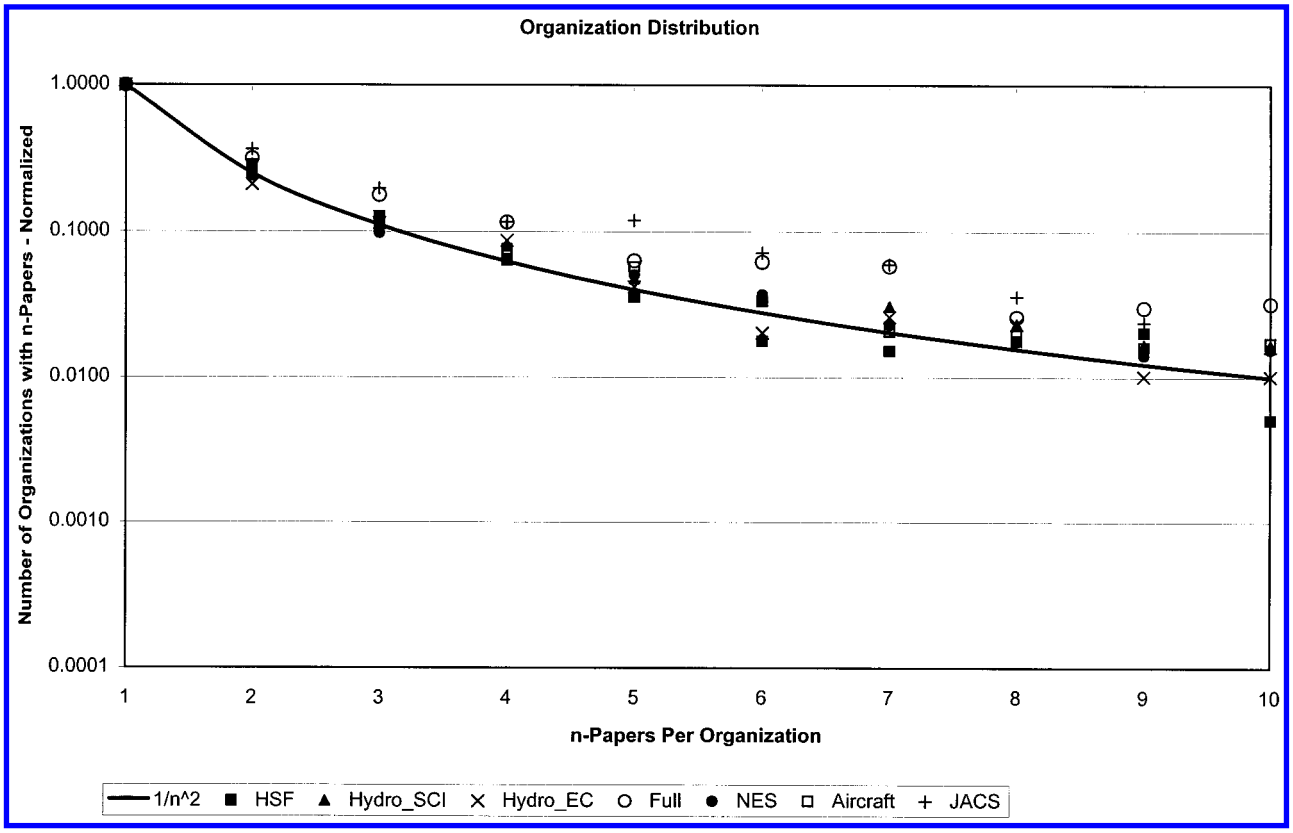


**Figure 3.**

in the database produced by a given institution, and the ordinate is the number of institutions that produced *n* relevant papers. In each case, the distribution function has been normalized to the number of institutions that produced one relevant paper.

The data center around a $1/n^2$ distribution remarkably well, although the FUL data exhibit the shallowest gradients again, for the same reasons as mentioned above. For a $1/n^2$ distribution, the number of organizations that generate three papers is about 11% of the organizations that generate one paper only. Also, integrating this distribution function shows that more than 67% of the papers result from organizations that produce three or less papers.

**4.1.4. Countries Producing Most Fullerene Papers.** There were 64 different countries listed in the SCI results. The dominance of a handful of countries was clearly evident (e.g., USA, 5861; JAPAN, 2840; GERMANY, 1500; PEOPLES R CHINA, 1363; RUSSIA, 1177; FRANCE, 1117; UK, 1001), but a series of small countries (SWITZERLAND, TAIWAN, BELGIUM, ISRAEL, SWEDEN, AUSTRIA, HUNGARY, THE NETHERLANDS) are also quite remarkably productive.

The UNITED STATES is more than twice as prolific as its nearest competitor (JAPAN) and is as prolific as its major competitors combined (JAPAN, GERMANY, PEOPLES REP OF CHINA). A 1997 study[14] listed the papers contributed by the top 50 nations to the world science literature, i.e., numbers of publications in the SCI. The top performers are in line with the bibliometric results of the seven DT studies.

**4.2. Citation Statistics on Authors, Papers, and Journals.** The second group of metrics presented is counts of citations to papers published by different entities. While citations are ordinarily used as impact or quality metrics,[15] much caution needs to be exercised in their frequency count interpretation, since there are numerous reasons why authors cite or do not cite particular papers.[16−18]

The citations in all the SCI papers were aggregated; the authors, specific papers, years, journals, and countries cited most frequently were identified and were presented in order of decreasing frequency. A small percentage of any of these categories received large numbers of citations. From the citation year results, the most recent papers tended to be the most highly cited. This reflected rapidly evolving fields of research.

**4.2.1. Most Cited Authors.** The citations in all 10 515 SCI papers were aggregated into a file of 263 844 entries, yielding an average of 25.1 references per paper. There were 33 579 different authors cited, with an average of 7.85 citations per cited author. A relatively few percent received large numbers of citations (e.g., KROTO HW, 4328; KRATSCHMER W, 3472; IIJIMA S, 1787; TAYLOR R, 1721; HADDON RC, 1711; HEBARD AF, 1563). However, in all the studies, the most cited authors, while prolific, are not the most prolific authors (except in one anomolous case, KROTO, in the FUL study), and vice versa. For example, the three most highly cited authors (KROTO HW, KRATSCHMER W, and IIJIMA S) ranked numbers 2, 36, 161, respectively, in the prolific authors list. The three most prolific authors (ACHIBA Y, KROTO HW, and KIKUCHI K) ranked numbers 197, 1, and 28, respectively, in citeability. Part of this difference may be due to the time lag between the highly cited authors' productivity at the time their highly cited papers were written and their productivity today, as well as the phase in their career of the prolific authors. Another partial explanation may be the intrinsic

**Table 5.** Cited Author Bibliometrics—SCI

| metric/study | FUL | JACS | NES | HYD | HSF | AIR | RIA |
|---|---|---|---|---|---|---|---|
| no. of papers retrieved | 10 515 | 2150 | 5481 | 4608 | 1284 | 4346 | 2300 |
| no. of citations | 263 844 | 85 000+ | 140 662 | 82 395 | 26 768 | 45 744 | 37 000+ |
| no. of citations per paper | 25.1 | 39.5 | 25.7 | 17.9 | 20.9 | 10.5 | 16.1 |
| no. of authors cited | 33 579 | 32 450 | 42 094 | 26 322 | 11 138 | 21 868 | 18 140 |
| no. of citations per author cited | 7.86 | 2.62 | 3.34 | 3.13 | 2.4 | 2.09 | 2 |
| no. of authors | 12 837 | 6535 | 12 453 | 7869 | 2483 | 6619 | 2975 |
| no. of citations per author | 20.6 | 13 | 11.3 | 10.5 | 10.8 | 6.9 | 12.4 |

nature of the papers; the large numbers of papers produced may reflect more applied papers, which lend themselves more to shorter-term production line type output. Stated differently, the time required to produce a fundamental seminal highly cited paper probably does not allow overly high volumes of papers to be produced.

Table 5 compares the bibliometric statistics for the different studies. Seven variables/figures of merit are presented for each study. The number of citations represents the total numbers of references in all papers retrieved. The average number of citations per paper is the ratio of the total number of citations to total number of papers retrieved. The number of authors cited is the total number of different first authors cited. The average number of citations per author cited is the ratio of the total number of citations to the total number of authors cited. The average number of citations per author is the ratio of references to authors.

From Table 5, there appears to be a difference between the more basic and applied areas in the average number of citations per paper. The more basic papers have more references than the applied papers. The basic papers tend to be more research-literature-oriented and are dependent on published documents, whereas the applied papers tend to be technology-product-oriented, with a reduced dependence on literature precedents and acknowledgments.

FUL clearly stands out in both average number of citations per author cited and average number of citations per author. FUL appears to be a young basic research field with a modest-sized core group of active researchers citing another modest-sized core group of active researchers, with much overlap between the two groups. Because the citations are focused on the modest-sized field of basic researchers, and not more broadly based as in the more mature technological fields, there is a substantial number of citations per author cited. Because of the breadth of research activity in FUL, paper authors are motivated to document this activity as extensively as possible. Both of these latter two metrics tend to decrease with increasing technical field maturity.

JACS is somewhat of an outlier to this trend in average number of citations per author cited. It should be remembered that JACS is far less focused than FUL, since JACS covers all of chemistry, and therefore would be expected to generate citations for a much broader group of authors than the more focused FUL. This dilution over many chemistry subdisciplines leads to less citations per author cited for JACS relative to FUL.

Figure 4 shows the distribution function of author citation frequency for the fullerene, NES, HSF, JACS, AIR, and HYDRO databases. The abcissa is the total number of citations $n$ received by a given author, and the ordinate is the number of authors that received $n$ total citations. In each case, the distribution function has been normalized to the number of authors that received one citation.

The data cluster very closely around a $1/n^2$ distribution, making this distribution far more universal than the somewhat discipline-dependent author publishing distribution. The FUL data are slightly above the curve and exhibit the shallowest gradients. This relationship between the FUL data and the other discipline data occurs in all the citation distribution functions and will be discussed in more detail in the next section on paper citation distributions.

Integration of this $1/n^2$ distribution function shows that over 67% of the citations are from authors cited three times or less. Some caveats are in order at this point. The citation data for Figures 4−6 represents citations generated only by the specific records in each database. It does not represent all the citations received by the references in those records; these references in the database records could have been cited additionally by papers in other technical disciplines. In addition, since very recent papers are included in the references, there is probably some skewing of the distribution function toward lower numbers of citations in these figures relative to distribution functions that do not include very recently published references. Recent papers do not have sufficient time to accumulate more than a small number of citations.

Conversely, the sample studies referenced in the next section do not have the two limitations described in the above paragraph. In the sample study, a small number of papers was selected. All citations to those papers from all fields were included, and a 4−5 year time interval between date of publication and the present was chosen to allow reasonable numbers of citations to accumulate.

**4.2.2. Most Cited Papers.** Table 6 compares the bibliometric statistics for the different studies. Four variables/figures of merit are presented for each study. The number of different papers cited is the total number of different papers referenced by the papers in the database. The average number of citations per cited paper is the ratio of the number of citations to the number of different papers cited. The average number of papers cited per author cited is the ratio of the total papers cited to the total authors cited.

There were 75 890 different papers cited, with an average of 3.48 citations per cited paper. Relatively few papers were highly cited (e.g., KRATSCHMER W 1990 NATURE V347, 2773; KROTO HW 1985 NATURE V318, 2319; HEBARD AF 1991 NATURE V350, 1177; IIJIMA S 1991 NATURE V354, 816). Relative to the other disciplines studied, the most highly cited FUL papers have larger numbers of citations (in some cases, orders of magnitude larger) and more recent publication dates. This reflects the more intensive FUL research activity and the young rapidly evolving nature of the field.

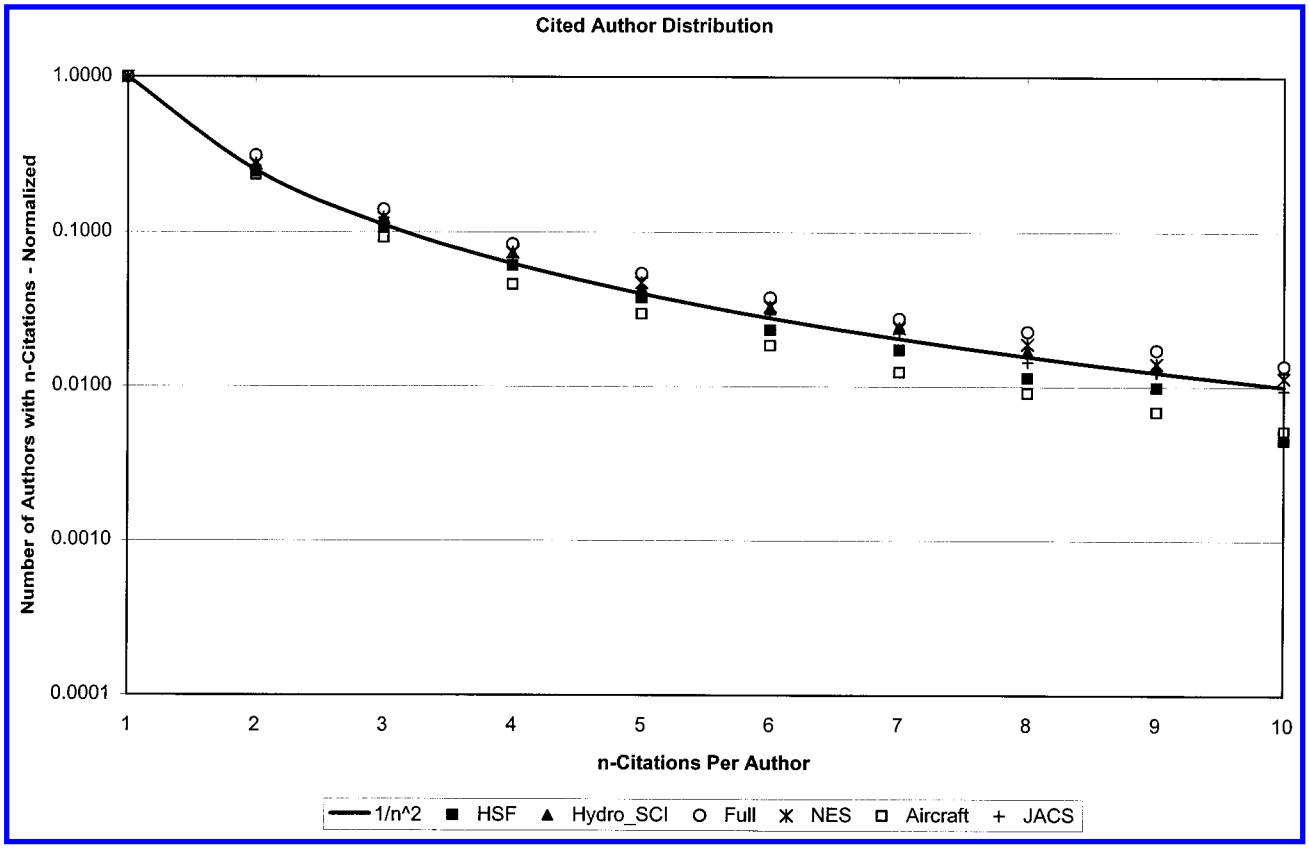


**Figure 4.**

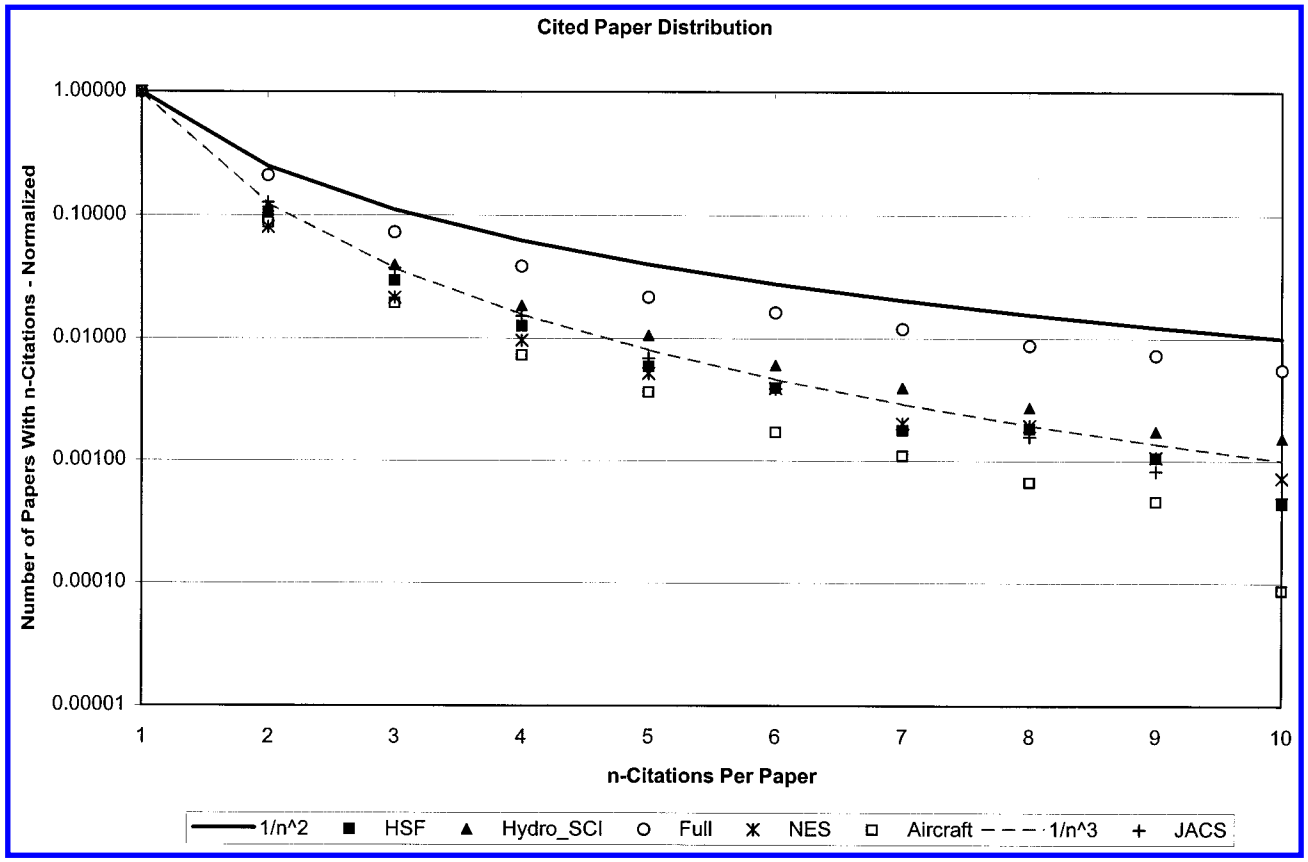


**Figure 5.**

From Table 6, there appears to be a trend in the average number of citations per cited paper, with this metric decreasing with increasing technical field maturity. This trend reflects the decreased dependence of the product-oriented applied papers on the research-oriented published literature, paralleling the conclusion reached in the previous section.

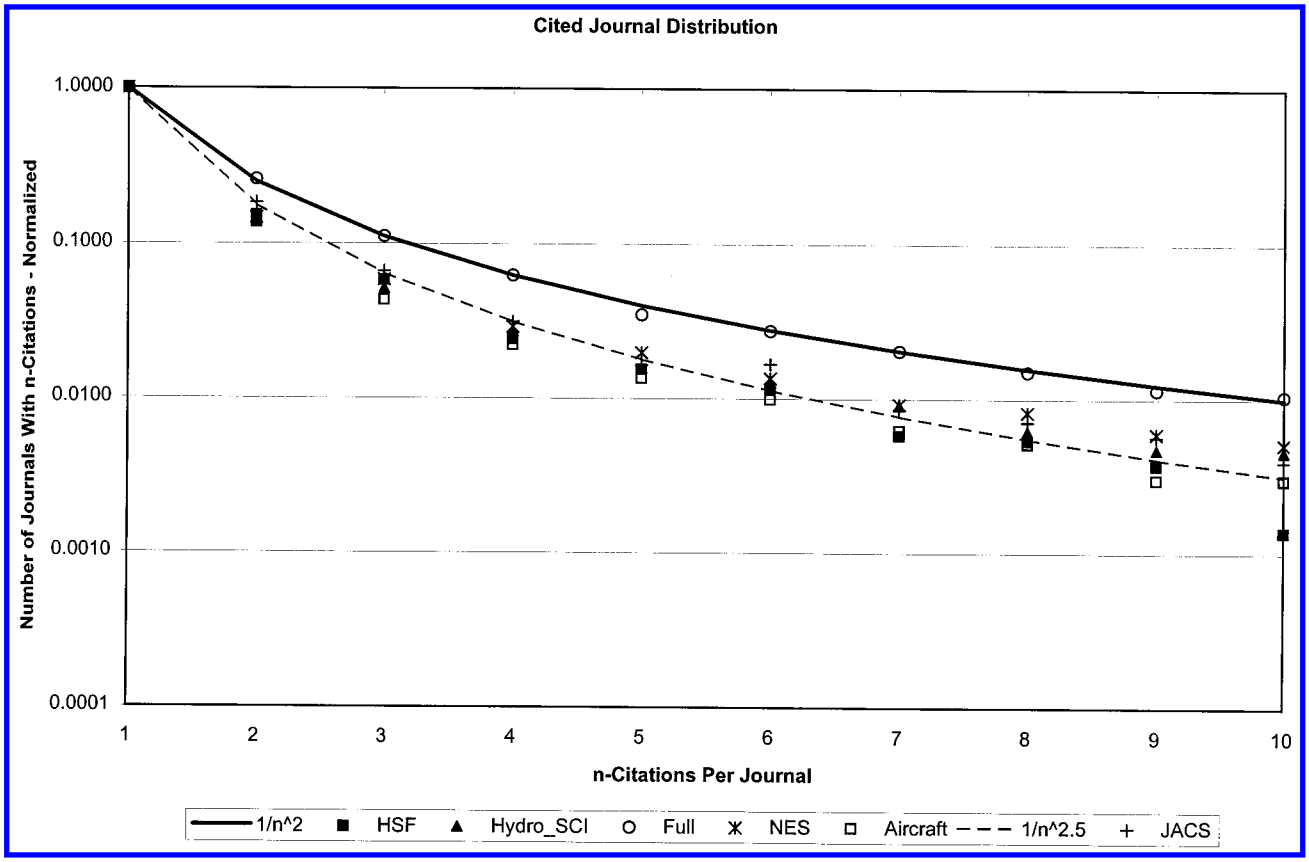


**Figure 6.**

**Table 6.** Cited Paper Bibliometrics—SCI

| metric/study | FUL | JACS | NES | HYD | HSF | AIR | RIA |
|---|---|---|---|---|---|---|---|
| no. of citations | 263 844 | 850 00+ | 140 662 | 82 395 | 26 768 | 45 744 | 370 00+ |
| no. of different papers cited | 75 890 | 64 800 | 93 194 | 57 618 | 20 950 | 38 792 | 30 400 |
| no. of citations per cited paper | 3.48 | 1.31 | 1.51 | 1.43 | 1.27 | 1.18 | 1.22 |
| no. of papers cited per author cited | 2.26 | 2 | 2.21 | 2.19 | 1.88 | 1.77 | 1.68 |

FUL stands out on this metric, again as a result of the concentration of the modest-sized community of citing researchers on the modest-sized community of active focused researchers.

**4.2.2.1. Aggregate Distribution Functions.** Figure 5 shows the distribution function of paper citation frequency for the fullerene, NES, HSF, JACS, AIR, and HYDRO databases. The abcissa is the total number of citations n received by a given paper, and the ordinate is the number of papers that received n total citations. In each case, the distribution function has been normalized to the number of papers that received one citation.

For five of the six topical fields presented, the data follow a $1/n^3$ distribution very closely, as contrasted with the $1/n^2$ distribution for author citations. Examination of the five topical studies that produced the five sets of data showed that each of the highly cited authors had a wide range of citations for his/her different papers. For any given highly cited author, most papers will receive few citations. It is the infusion of numbers of lowly cited papers from the highly cited authors which expands the pool of lowly cited papers in Figure 5 and results in the conversion of the $1/n^2$ distribution of Figure 4 to the $1/n^3$ distribution of Figure 5. This effect appears to transcend the five different science and technology topical fields and to be almost universal based on the limited data presented for the six topical science and technology fields. The resulting relation among the distribution functions, the Kostoff−Eberhart−Toothman (KET) law,[6] can be restated as follows: for a topical science and technology field, the ratio of the normalized number of authors with $n$ citations per author to the normalized number of papers with $n$ citations per paper is $n$, for low to moderate values of $n$.

The FUL distribution from Figure 5 is between a $1/n^3$ and $1/n^2$ distribution. Its apparent modest deviation from the KET law prediction, however, is somewhat muted by the FUL author distribution from Figure 4, also lying slightly above the $1/n^2$ average of the other five disciplines. In Figure 5, the AIR distribution function exhibits the highest gradient, and the FUL distribution function exhibits the lowest one. The differences between these two distributions reflect the intrinsic differences of the maturity of the underlying disciplines. Aircraft S&T has been an established topical area for many years. The technology/science ratio is perhaps the highest of all the six disciplines studied. Fullerenes were discovered in the mid-1980s. As the DT analyses will show in the later sections of this paper, fullerenes S&T are essentially at the basic research experimentally-focused stage, based on the published journal literature. Its technology/

**Table 7.** Cited Journal Bibliometrics—SCI

| metric/study | FUL | JACS | NES | HYD | HSF | AIR | RIA |
|---|---|---|---|---|---|---|---|
| no. of citations | 263 844 | 850 00+ | 140 662 | 82 395 | 26 768 | 45 744 | 370 00+ |
| no. of different journals/sources cited | 13 294 | 6725 | 28 740 | 21 523 | 9498 | 21 518 | 2975 |
| no. of citations per cited journal | 19.85 | 12.6 | 4.89 | 3.83 | 2.82 | 2.13 | 0.00 |
| no. of authors | 12 837 | 6535 | 12 453 | 7869 | 2483 | 6619 | 18 140 |
| no. of journals cited per author | 1.04 | 1.03 | 2.31 | 2.74 | 3.83 | 3.25 | |
| no. of authors cited | 33 579 | 32 450 | 42 094 | 26 322 | 11 138 | 21 868 | |
| no. of authors cited per journal cited | 2.53 | 4.83 | 1.46 | 1.22 | 1.17 | 1.02 | |

science ratio is the lowest of the six disciplines studied. The other five disciplines have established an equilibrium between science and technology, whereas fullerenes are still following a start-up transient toward this equilibrium.

As shown in recent S&T data mining studies,[6,7] the more basic papers tend to receive more citations than the applied papers, and the more basic journals consequently receive more citations than the more applied journals. Thus, in an S&T field such as aircraft, which has a substantial ratio of applied to basic papers, there are fewer papers that are realistic candidates for a high number of citations. The ratio of aircraft papers that receive a large number of citations to those receiving one citation would therefore be relatively small. Conversely, in an S&T field such as fullerenes, which has a small ratio of applied to basic papers, there are many more papers that are realistic candidates for a high number of citations. The ratio of fullerene papers that receive a large number of citations to those that receive one citation would therefore be relatively large compared to aircraft. The data support this argument, and if/when fullerenes will advance into the technology development stage from the published literature perspective, the fullerene distribution function of Figure 5 would be expected to evolve to the distribution function predicted by the KET law. In some sense, the KET law can be viewed as a metric of the basic/applied balance, or equilibrated developmental maturity, of an S&T discipline.

**4.2.2.2. Characteristics of Highly Cited and Poorly cited Papers.** To ascertain whether any relationship between highly cited and lowly cited papers and their associated journals and performing organizations could be observed, the characteristics of samples of highly and lowly cited papers were analyzed. Two small sample analyses led to a conclusion for hypersonic flow[6] that Russia produced slightly more papers than the leading organizational producer (NASA) but had almost no highly cited papers. Whether these differences extend beyond supersonic/hypersonic flow to other topical areas is an interesting question.

The systemic difference between Russian and American papers from the hypersonics study led to a small focused study on a subset of fullerenes.[19] All English language papers in the SCI published in 1993/1994 that contained the phrase CARBON NANOTUBE* were selected. There were 131 such papers; all were relevant to the desired topic. The citation patterns of papers written by Russian authors only and American authors only were examined.

There were 44 papers published by American authors only, and three papers by Russian authors only. The American papers averaged 27.3 cites per paper, while the Russian papers averaged 6 cites per paper. The American median was 20 cites per paper, while the Russian median was 4 cites per paper. (As an aside, the Japanese papers appeared to very numerous and well-cited, followed by the Western European papers).

Whether or not the Russians are prolific in a field in terms of paper production, their works are not getting cited by the larger technical community. Possible explanations are:

(1) They could be doing good (citeable) work and not reporting it.

(2) The work reported may be good, but very applied, and not amenable to citing in the literature; i.e., citation is not the appropriate measure of quality or utility or impact in this case.

(3) The work reported could be good, but might not be published in the forefront literature, and the technical community therefore might not be very aware of this work.

(4) The work could be poor and the citations poor.

There are two crucial pieces of data missing from these short studies (and from most bibliometrics analyses) that prevent harder conclusions about quality and value to be drawn. The amount of research effort represented by each paper is unknown to the analyst, and the eventual use of the results from each paper is unknown to the analyst. Thus, the number of highly cited papers per dollar of research investment (or some similar research efficiency metric), probably a better measure of value than pure numbers of papers or highly cited papers, cannot be stated. In the hypersonics case, the quality of the eventual hypersonic vehicles that resulted from the papers' research, probably a better measure than numbers of cited papers, was not tracked and cannot be stated. In addition, the papers in the two short hypersonic studies were not read in detail independently by hypersonic flow experts, and thus their quality could not be gauged independently from another perspective and correlated to the citation results.

**4.2.3. Most Cited Journals.** There were 13 294 different journals and other sources cited. Relatively few sources were highly cited (e.g., NATURE, 21 773; CHEM PHYS LETT, 20 735; J AM CHEM SOC, 19 534; PHYS REV B, 17 985; PHYS REV LETT, 15 482; J PHYS CHEM US, 15 120; SCIENCE, 11 801).

Table 7 compares the bibliometric statistics for the different studies. Seven variables/figures of merit are presented for each study. The number of different journals/sources cited is the total number of different journals and other sources referenced by the papers in the database. The average number of citations per cited journal is the ratio of the number of citations to the number of different journals and other sources cited. The average number of journals cited per author is the ratio of total journals and other sources cited to total authors. The average number of authors cited per journal cited is the total number of authors cited to the total number of journals and other sources cited.

Fullerenes is the most basic of the six S&T areas studied with DT so far, based on the journal publications literature. It has the strongest journal correlation between high numbers of publications and citations. In the previous DT studies, some journals tended to publish many topical papers and be highly cited, some journals tended to publish many topical papers but not be highly cited, and some journals tended to publish relatively few topical papers but be highly cited. Most of the disciplines studied had a technology component along with a research component. The topical published papers tended to be slightly more applied than some of their references, and thus the journals which contained a large number of the topical published papers tended to be more applied than the journals which contained their more basic references. These more basic journals tended to rank higher in citations relative to publications, while the more applied journals tended to rank higher in publications relative to citations. Fullerenes is a relatively young topical area, and the bulk of the S&T effort is concentrated on research. Most of the papers are basic research, and the thrust of most of the journals that publish these papers is also basic.

There is a definite trend in average number of citations per cited journal, decreasing sharply from the basic fields to the applied fields. One needs to make a distinction here between the journals in which authors publish and the journals that they cite.

As the Bradford's law results showed, there were more credible journals in which the researchers could publish in the basic fields compared to the applied fields. However, in the case of citations, there is a wider variety of journals that the researchers in the applied fields will access (both basic and applied journals) than the researchers in the basic fields will access (basic). Therefore, it would be expected that the researchers in basic fields (who cite more frequently as shown above, and who cite a narrower group of journals than their applied counterparts) would have a substantially higher value of this "citations per cited journal" metric than their applied counterparts.

This difference in breadth of journals cited between the researchers in basic and applied fields, discussed in the previous paragraph, is substantiated and displayed most dramatically by the average number of journals cited per author metric. The metric increases sharply from the basic fields to the applied fields.

The final metric listed, average number of authors cited per journal cited, trends downward as the fields become more applied, with the lone exception of JACS. As stated previously, the researchers in the more applied fields tend to cite from a wider variety of journals than their counterparts in the more basic fields, and the denominator of this metric therefore increases as the fields become more applied. In the JACS case, the number of authors cited is slightly exaggerated because of its breadth of coverage, as shown in Table 6. This effect would tend to increase the metric numerator modestly. Probably the more pronounced effect derives from the tendency of authors in a given journal to cite that journal more frequently than would be expected on average. Since JACS was the only study in which a single journal was used, there is probably some skewing of the JACS authors toward citing JACS papers, and hence the anomalous value of the final metric.

Figure 6 shows the distribution function of journal citation frequency for the fullerene, NES, HSF, JACS, AIR, and HYDRO databases. The abcissa is the total number of citations $n$ received by a given journal, and the ordinate is the number of journals that received $n$ total citations. In each case, the distribution function has been normalized to the number of journals that received one citation.

The data follow approximately a $1/n^{2.5}$ distribution. Paralleling the distributions of Figure 5, FUL exhibits the shallowest gradient and AIR exhibits the steepest one. The reasons for these differences are identically those behind the Figure 5 differences and need not be repeated here.

As Bradford's law suggests, there is a concentration of papers in the higher-quality core journals. When this is coupled with the strong nonlinearity of the distribution of cited papers as shown in the previous section, a further separation among journals (than the $1/n^2$ average distribution of Figure 2) based on citations received would be expected. This effect is strongly muted because the wide disparity in citations per paper within a given journal is integrated out to arrive at the citations per journal for all papers published by the journal.

The authors end this bibliometrics section by recommending that the reader interested in researching the topical field of interest would be well-advised to, first, obtain the highly cited papers listed and, second, peruse those sources that are highly cited and/or which contain large numbers of recently published topical area papers.

**4.3. Database Tomography Results. 4.3.1. Most Frequently Used Keywords in Science Citation Index.** The frequency distributions of the SCI keywords associated with each paper were analyzed. An overall picture of fullerene S&T emerges. The field may be divided into three categories, listed in descending levels of effort: materials, experiments, and theory/computation.

Overall, the field appears exclusively focused on basic research, with the majority emphasis on experiments. There are essentially no technology terms, no application terms, and very few terms relating to computation and modeling. The major focus is on C60 and C70, isomers, and polymers, although some of the higher fullerenes are alluded to as well. Some alternative sources are described, such as soot and flames. There are relatively few indications that elements other than carbon-based are being examined for ordered fullerene-like structures.

The major geometries examined are closed cages and thin-walled variants (tubules, thin films, and shells), and the major phases/structures revolve around clusters, single crystals, and gas and solid phases. A substantial amount of the research appears centered around the molecular, atomic, and electronic structures. Crystal lattice structures are examined heavily, as are electronic states and their associated resonances and vibrational frequencies.

While the majority of experimental activity centers on molecular and electronic spectra, there appears to be some effort on macrochemistry approaches. The instrumental techniques emphasize diffraction and scattering over a broad radiation spectrum. The phenomena studied focus on growth, phase transition and separation, emission over a wide frequency spectrum, and alignment and ordering of the structure at the molecular level.

The taxonomy derived from the keywords, along with sample keywords in each category, is as follows:

materials

*macroclasses:* COMPLEXES, 209; CONDUCTING POLYMER, 71; POLYCYCLIC AROMATIC−HYDROCARBONS, 62

*macromaterials/-compounds*: C60, 5770; FULLERENES, 1907; DIAMOND, 107

*macrodescriptors*: STABILITY, 194; MODES, 102; SYSTEMS, 93; INHIBITION, 54

*macroproperties/-parameters*: SUPERCONDUCTIVITY, 538; PHOTO−PHYSICAL PROPERTIES, 171; TEMPERATURE, 124; PHOTOCONDUCTIVITY, 97

*macrophases/-structures*: SOLID C60, 703; CARBON CLUSTERS, 244; SINGLE-CRYSTAL C60, 68

*macrogeometries*: FILMS, 473; CARBON NANOTUBES, 246; MICROTUBULES, 237; SPHEROIDAL CARBON SHELLS, 95

*microdescriptors*: MOLECULES, 464; IONS, 218; ATOMS, 139; CARBON CLUSTER IONS, 82

*microphases/-structures*: ELECTRONIC STRUCTURE, 337; CRYSTAL STRUCTURE, 130; GROUND STATE, 68

experiments

*instrumental tools*: SPECTROSCOPY, 381; SCANNING TUNNELING MICROSCOPY, 117; MASS SPECTROMETRY, 96; CYCLIC VOLTAMMETRY, 58; PULSE RADIOLYSIS, 58

*interaction phenomena*: ABSORPTION, 160; DIFFRACTION, 50; RAMAN SCATTERING, 49; NEUTRON SCATTERING, 43

*interaction-induced changes*: GROWTH, 339; PHASE TRANSITION, 94; PHOTOINDUCED ELECTRON TRANSFER, 86

*instrumental output*: SPECTRA, 509; ABSORPTION SPECTRA, 147; SPECTRUM, 74

theory/computation

MOLECULAR DYNAMICS, 182; MODEL, 175; APPROXIMATION, 46; ABINITIO, 41

**4.3.2. Phrase Frequency Analysis−Pervasive Themes in Science Citation Index.** High-frequency single, double, and triple word phrases from the text of the SCI database whose technical contents were deemed by topical experts to be significant were identified as the pervasive themes. Nontechnical content phrases, trivial phrases (automatically), etc., were eliminated from the analysis. In this particular exercise, the database was split into two parts, titles and abstracts, and the analysis was done on each part. Since the highest frequency phrases from the title and abstract databases were very similar, only raw data outputs from the abstract database will be considered here.

The significant high-frequency phrases were contextually integrated to form the following coherent picture of the main database structural elements. From a global perspective, the SCI fullerene database has three components: two are major and one is minor. Parallel to the keyword taxonomy, the major abstract categories are materials and experiments, and the minor one is theory/computation. There are many subcategorizations possible within these category headings; one useful taxonomy is the following:

**4.3.2.1. Materials.** This category is classified into the following:

*materials classes*: METAL, 829; COMPOUNDS, 786; POLYCYCLIC AROMATIC HYDROCARBONS, 67

*materials and compound types*: FULLERENE(S), 6791; CARBON, 3574; GRAPHITE, 1065

*macro level descriptors*: C[−]60, 12 879; C[−]70, 2481; CARBON CLUSTER(S), 362; SINGLE CRYSTAL(S), 371

*microlevel descriptors*: CARBON CLUSTER IONS, 45; STRUCTURE OF C[−]60, 90; PENTAGONAL PINCH MODE, 29

*geometries*: FILM(S), 2518; SURFACE, 1354; SINGLE-WALL CARBON NANOTUBES, 30

*macrophases and macrostructures*: STRUCTUR*, 4739; PHASE, 1760; ELECTRONIC STRUCTURE(S), 528; ELECTRONIC PROPERTIES, 189

*microphases and microstructures*: ELECTRON*, 4466; MOLECUL*, 5013; CARBON ATOMS, 291; EXCITED SINGLET STATE, 31

*critical environmental parameters*: TEMPERATURE(S), 3105; DENSITY, 945; PRESSURE, 689; SUPERCONDUCTING TRANSITION TEMPERATURE, 62

**4.3.2.2. Experiments.** This category is classified into the following:

*experimental activity*: OBSERVED, 2339; EXPERIMENTAL, 1021; MEASURE*, 1769; INVESTIGATED, 847

*instrumental tools*: LASER, 840; ELECTRON MICROSCOPY, 352; MASS SPECTROMETRY, 296; SCANNING TUNNELING MICROSCOPY, 115

*interaction phenomena*: ABSORPTION, 1148; X-RAY DIFFRACTION, 292; RAMAN SCATTERING, 99

*interaction-induced changes*: TRANSITION, 1165; CHARGE TRANSFER, 392; 122, LASER DESORPTION, 122

*instrumental measurements*: SPECTR*, 2795; EXPERIMENTAL DATA; 180, EXPERIMENTAL RESULTS; 179, ABSORPTION SPECTRA, 179

**4.3.2.3. Theory/Computation**: (MODEL, 1381; AB INITIO CALCULATIONS, 66; MOLECULAR DYNAMICS

SIMULATIONS, 59; DENSITY FUNCTIONAL THEORY, 44). The main focus of the research is clearly experimental, with a much smaller effort devoted to theory and computational approaches. There does not appear to be a diversity of computational approaches; the focus is on HOMO/LUMO molecular orbital calculations.

The instrumental focus is on diverse aspects of spectroscopy and microscopy at the atomic/molecular level. Main areas of study are the diverse phase and state transitions due to the effects of incoming radiation, especially in the electron state changes. Mass and radiation spectra are the major types of outputs studied. Experiments at the macrolevel, such as strength or fatigue tests, or chemical thermodynamics tests or techniques, are not listed.

Main material types studied are the metal compounds, with prime focus on C60 and C70, but with some attention paid to the higher fullerenes and fullerene derivatives. Macrostates studied most extensively are clusters, polycrystals, and single crystals, mainly in thin-film and nanotube geometries. The macrolevel properties studied center around electronic primarily and optical secondarily. There is essentially negligible effort on studying mechanical and chemical/thermodynamic properties. The microstates studied center around electronic energy and transition states, at the atomic/molecular/ionic level. This couples well with the instrumental focus on spectroscopy and microscopy.

Noticeable by their absence are any mention of technology or application issues in the particular published journal literatures accessed here. After 15 years since fullerenes were discovered, and substantial promises about new applications were made, the field as represented by the published journal literature remains locked in basic research at the microstructure level. [*A short query of the IBM patent database (fullerene* OR carbon nanotube* OR C60 OR C-60 OR C70 OR C-70) retrieved about 700 patents. If patents are perceived as the gateway to applications, then there may be some movement of the fullerene topical area toward applications, but it is not being reflected in the published journal literature*.] Very limited geometrical configurations are mentioned in the published journal literature results, mainly focused on thin structures, and there appear to be no forays to explore similar phenomena in other materials beside carbon.

This analytical procedure, and subsequent analytical procedures based on the phrase proximity results (described later), are not independent of the analyst's domain knowledge; they are, in fact, expert-centric. The computer techniques play a strong supporting role, but they are subservient to the expert, and not vice versa. The computer-derived results help guide and structure the expert's analytical processes; the computer output provides a framework upon which the expert can construct a comprehensive story. The conclusions, however, will reflect the biases and limitations of the expert(s). A fully credible analysis requires not only domain knowledge by the analyst(s) but probably domain knowledge representing a diversity of backgrounds.

**4.3.3. Phrase Proximity Analysis—Relationships among Themes. 4.3.3.1. Background.** To obtain the theme and subtheme relationships, a phrase proximity analysis is performed about each theme phrase. Typically, 40−60 multiword phrase themes are selected from a multiword phrase analysis of the type shown above. For each theme phrase, the frequencies of phrases within ±50 words of the theme phrase are computed for every occurrence of the theme phrase in the full text. A phrase frequency dictionary is constructed that contains the phrases closely related to the theme phrase. Numerical indices are employed to quantify the strength of this relationship. Both quantitative and qualitative analyses of each phrase frequency dictionary (hereafter called cluster) yield those subthemes closely related to the main cluster theme.

Then, threshold values are assigned to the numerical indices. These indices are used to filter out the cluster member phrases most closely related to the cluster theme. For purposes of analysis, the cluster members in a given theme are segregated by their values of inclusion indices Ii and Ij. Ii is the ratio of Cij (the frequency with which the cluster member phrase occurs within ±*M* words of the theme phrase) to Ci (the total text cluster member phrase frequency) and is the inclusion index based on the cluster member. Ij is the ratio of Cij to Cj (the total text cluster theme phrase frequency) and is the inclusion index based on the theme phrase. Ii and Ij are categorized as either high or low. The dividing points between high and low Ii and Ij are the middle of the "knee" of the distribution functions of numbers of cluster members vs values of Ii and Ij. All cluster members with Ii greater than or equal to approximately 0.5 were defined as having high Ii. All cluster members with Ij greater than or equal to 0.1 were defined as having high Ij.

A high value of Ii means that whenever the cluster member appears in the total database text, there is a high probability that the theme phrase will appear within ±50 words of the cluster member. A high value of Ij means that, whenever the theme phrase appears in the total database text, there is a high probability that the cluster member will appear within ±50 words of the theme phrase.

Thus, phrases categorized as high Ii high Ij are coupled very strongly to the theme phrase. Whenever the theme phrase appears in the total database, there is a high probability that the cluster member will be physically close. Whenever the cluster member appears in the total database, there is a high probability that the theme phrase will be physically close. Whenever either phrase appears in the total database text, the other will be physically close.

Consider phrases categorized as low Ii high Ij. Whenever the cluster member appears in the total database text, there is a low probability that it will be physically close to the theme phrase. Whenever the theme phrase appears in the total database text, there is a high probability that it will be physically close to the cluster member. This type of situation occurs when the frequency of occurrence of the cluster member Ci is substantially larger than the frequency of occurrence of the theme phrase Cj, and the cluster member and the theme phrase have some related meaning.

Single word phrases have absolute frequencies an order of magnitude higher than double word phrases. Thus, the phrases categorized as low Ii high Ij are typically high-frequency single word phrases. They are related to the theme phrase but much broader in meaning than the theme phrase. A small fraction of the time that these broad single word phrases appear, the more narrowly defined double word phrase theme will appear physically close. However, whenever the narrowly defined double word phrase theme appears, the broader related single word phrase cluster member will

appear. The phrases under this heading can also be viewed as a higher level taxonomy of technical disciplines related to the theme.

Consider phrases categorized as high Ii low Ij. Whenever the cluster member appears in the total database text, there is a high probability that it will be physically close to the theme phrase. Whenever the theme phrase appears in the total database text, there is a low probability that it will be physically close to the cluster member. This type of situation occurs when the frequency of occurrence of the cluster member Ci is substantially smaller than the frequency of occurrence of the theme phrase Cj, and the cluster member and the theme phrase have some related meaning. Thus, the phrases categorized as high Ii low Ij tend to be low-frequency double and triple word phrases, related to the theme phrase but very narrowly defined.

A large fraction of the time that these very narrow double and triple word phrases appear, the relatively broader double word phrase theme will appear physically close. However, a small fraction of the time that the relatively broad double word phrase theme appears, the more narrow double and triple word phrase cluster member will appear. This grouping has the potential for identifying "needle-in-a-haystack" type thrusts that occur infrequently but strongly support the theme when they do occur. For studies whose main focus is innovation and discovery through related literatures,[10] this grouping would be central in linking strongly related literatures. One of many advantages of full text over key or index phrases is this illustrated ability to retain low-frequency but highly important phrases, since the key phrase approach ignores the low-frequency phrases.

Finally, consider phrases categorized as low Ii low Ij. Whenever the cluster member appears in the total database text, there is a low probability that it will be physically close to the theme phrase. Whenever the theme phrase appears in the total database text, there is a low probability that it will be physically close to the cluster member. The phrases in this category are intermediate to the high-technical content, low-frequency high Ii low Ij typically triple word phrases and the low-technical content, high-frequency low Ii high Ij typically single word phrases. Thus, the low Ii low Ij phrases tend to have moderate technical content, moderate frequency, and typically double word structure. These phrases tend to be most useful in providing a balanced technical description of the relationships within the cluster of interest; they are not too detailed to obscure the larger context, yet they are not too general to be devoid of meaningful content.

**4.3.3.2. Analysis.** The full text database was split into two databases. One was the abstract narrative database (referred to as ABSTRACT in the phrase proximity analysis below), and phrase proximity analysis of this database yielded mainly topical theme relationships. The other database (referred to as BLOCK below) consisted of records (one for each published paper) containing four fields: author(s), title, journal name, and author(s) institutional address(es). Phrase proximity analysis of this database yielded not only topical theme relationships from the proximal title words but also relationships between themes and authors, journals, and institutions.

Because of space limitations in this document, only one theme, CARBON NANOTUBES, was chosen for the phrase proximity analysis. It was high frequency in both the abstracts and titles and is a central theme in recent fullerene research. In the following section, the cluster theme CARBON NANOTUBES is analyzed for the BLOCK and ABSTRACT database components. Further, for each of these database components, the cluster theme is analyzed from the two perspectives of high Ii low Ij and low Ii high Ij. The phrase proximity analysis process for CARBON NANOTUBES consisted of providing the experts with two lists of cluster members, one sorted by Ii and the other by Ij. By visual examination of these lists, the experts constructed categories of related items, and these relationships are reported below.

**4.3.3.3. Phrase Proximity Analysis—CARBON NANOTUBES. 4.3.3.3.1. BLOCK Database; Low Ii High Ij.** The phrases describe the more generic associations with CARBON NANOTUBES. The journal and departmental emphases are interdisciplinary and relate to physics, chemistry, and materials. The major performing institutions are dispersed throughout the world. In modest contrast to the total database, Japan appears to have greater efforts than the USA in this area. In terms of sheer numbers, the Asian countries are matching the western democracies in output. Geometries focus on thin-walled tubes, experiments focus on growth and nanoscale microscopy, and computational efforts appear minimal.

The taxonomy categories and sample entries are as follows:

authors: ANDO, 31; DRESSELHAUS, 31; SMALLEY, 27

journals: PHYSICAL REVIEW [LETTERS], 98; [APPLIED] PHYSICS LETTERS, 92; [PHYSICAL REVIEW] B-CONDENSED MATTER, 71; JOURNAL OF CHEMICAL PHYSICS, 64; CHEMICAL PHYSICS LETTERS, 56; NATURE, 35

institutions: BERKELEY, 128; RICE, 105; OXFORD, 76; TSUKUBA, 63; NEC, 55; IBARAKI, 47; MIE, 44; INDIAN INST SCI, 40

departments: PHYS, 499; CHEM, 292; MAT, 236

states/provinces/cities: CA, 106; TOKYO, 87; BEIJING, 73; TX, 67; MA, 65; LAUSANNE; 64; SUSSEX, 48; BANGALORE, 44; KARNATAKA, 42

countries: JAPAN, 286; USA, 176; CHINA, 102; ENGLAND, 93; FRANCE, 91

materials: CARBON, 228; GRAPHITE, 98; SOLID, 77

geometries/dimensions: TUBULES, 227; MICROTUBULES, 160; GRAPHENE TUBULES; SINGLE-WALL, 32

experiments: GROWTH, 203; MICROSCOPY, 84; SYNTHESIS, 55

particles: ELECTRON, 91; ELECTRONIC STRUCTURE, 46; NANOPARTICLES, 39; PARTICLES, 35

computational: MODEL, 56

**4.3.3.3.2. BLOCK Database; High Ii Low Ij.** The phrases describe the more specific associations with CARBON NANOTUBES. A large number of fullerene researchers appear fully centered on carbon nanotubes, although only one journal (*Physical Review B*) appeared to have such a fullerene subarea focus.

The taxonomy categories and sample entries are as follows:

authors: AJIKI, RINZLER, SMALLEY, SALVETAT, NAGY, BERNAERTS

journals: PHYSICAL REVIEW B-CONDENSED MATTER

institutions: CSIC INST CARBOQUIM, BANGALORE, CORNELL, RES CTR MOL MAT OKAZAKI AICHI, MATTER NATL TSING, OFF NATL RECH AEROSP PHYS, FLORIDA ATLANTIC UNIV

departments: DEPT PHYS, CHEM ENG, CTR MOL MAT, SINICAL INST MET

states/provinces/cities: HI, CHIKUSA, CHATILLON, LAUSANNE, TOYAMA, BEIJING, LOUVAIN, AZ, DELFT, NC

countries: ENGLAND, JAPAN, FRANCE, USA, SWITZERLAND, TAIWAN, THE NETHERLANDS, BELGIUM

materials: POLYACETYLENE MICROTUBULES, GRAPHENE TUBULES, METALS, LANTHIUM CARBIDE, BORON NITRIDE

geometries/dimensions: MULTIWALL, MICROTUBULES RINGS, CYLINDER, QUANTUM WIRES

phenomena: AHARONOV−BOHM EFFECT, PERSISTENT CURRENTS, CATALYTIC SYNTHESIS

experiments: ARC DISCHARGE, NANODIFFRACTION, GROWTH TUBES, VAPOR-GROWN, SCANNING ELECTRON, FIELD EMISSION

particles: NANOSCALE GRAPHITIC, NANOPARTICLES, PLASMONS

computational: MODEL TUBE, MOLECULAR SIMULATION

**4.3.3.3.3. ABSTRACT Database; Low Ii High Ij.** The phrases describe the more generic associations with CARBON NANOTUBES. The main material classes examined are metals and composites, and the main carbon class examined is graphite. An application focus appears to be catalysis-related. The experimental focus is on nanoscale microscopy and diffraction, with some electrochemical discharge-oriented research as well. The major experimental purpose appears to be studying growth and deposition and some emphasis on decomposition.

The taxonomy categories and sample entries are as follows:

materials

macroclasses: METAL, 57; MATERIALS, 42; COMPOSITE, 27

macromaterials/compounds: CARBON, 319; GRAPHITE, 119; NI, 31; GRAPHENE, 26

macrodescriptors: CATALYST, 36; CATALYTIC, 30; CATALYSTS, 24

macroproperties/parameters: PROPERTIES, 102; DIAMETER, 71; NM, 67; MAGNETIC, 59; DEGREES, 54; CONDUCTANCE, 22

macrophases/-structures: STRUCTURE, 114; DEFECTS, 29; ELECTRICAL, 26

macrogeometries: NANOTUBES, 372; SINGLE-WALL, 47; SHEETS, 25; BUNDLES, 24

microproperties: ALIGNED, 28; AXIS, 28; PERPENDICULAR, 23; TIGHT-BINDING, 23

microphases/structures: ELECTRON, 171; PARTICLES, 42; ATOMIC, 39; NANOPARTICLES, 37; PLASMON, 20

experiments

instrumental tools: TRANSMISSION ELECTRON MICROSCOPY, 53; SCANNING, 44; CATHODE, 28; STM, 25

interaction phenomena: DIFFRACTION, 49; DISCHARGE, 27

interaction-induced changes: GROWTH, 78; DECOMPOSITION, 21; DEPOSIT, 20

detector measurements: IMAGES, 39; HIGH RESOLUTION, 29

theory/computation: MODEL, 62

**4.3.3.3.4. ABSTRACT Database; High Ii Low Ij.** The phrases describe the more specific associations with CARBON NANOTUBES. Spatial variations of microproperties both along the tubes (e.g., LDOS near the tip structures) and perpendicular to the tubes' axis (e.g., DMS in the radial direction), and their impact on transport and emission properties, seem to be of specific interest. Also, lattice instabilities related to distortion of Kekule structures by these magnetic fields perpendicular to the tubes' axis seem to be of specific nanotube interest.

The taxonomy categories and sample entries are as follows:

materials

macroclasses: METALLIC FIBERS, HYDROGEN REDUCED CATALYSTS, COBALT

macromaterials/-compounds: CARBON DEPOSIT, YC2, POLYCRYSTALLENE GOLD

macroproperties/-parameters: CH4 GAS PRESSURE, CAPILLARITY, TENSILE

macrophases/-structures: SEAMLESS, SINGLE CRYSTALS ENCAPSULATED, MOLTEN

macrogeometries: ROPES OF SINGLE-WALLED, BUNDLES OR ROPES, KEKULE AND OUT-OF-PLANE, MULTIWALL

microdescriptors: PI SIGMA PLASMON

microproperties: LDOS, DMS

microphases/-structures: LATTICE INSTABILITY

microgeometries: MAGNETIC FIELD PERPENDICULAR

experiments

tools: BEAM OF DIAMETER, DC ARC-DISCHARGE, GRAPHITE CATHODE, ELECTRON MICROSCOPY HREM

interaction phenomena: PYROLYSIS OF BENZENE, MEANS OF TRANSMISSION, ACT AS NUCLEATION, NANODIFFRACTION, MAGNETIC FLUX

interaction-induced changes: PURIFIED NANOTUBES, DECOMPOSITION OF ACETYLENE, FIELD EMISSION, GASIFICATION, OPENED

detector measurements: OBSERVED BY SCANNING, PATTERNS OBTAINED, HRTEM IMAGES

theory/computation: LOW-ENERGY THEORY, MODEL IS DEMONSTRATED, THEORY FOR SINGLE-WALL, REALISTIC MANY-BODY POTENTIAL, AHARANOV−BOHM

## 5. LESSONS LEARNED

When applying the phrase frequency algorithm to a database, it is valuable to examine different record fields. There was little overall difference in conclusions about fullerenes from the keyword or abstract phrase frequency analysis, probably due to its relatively early research stage. However, the initial application of the phrase frequency analysis to the aircraft database was to integrate specific category results into a global perspective or mosaic of aircraft S&T. The analysis was applied to the keyword field and then to the abstract field. While many of the category judgments were the same from the different field perspectives, some were very different. For example, the keywords portrayed aircraft S&T as longevity and maintenance oriented, with minimal effort on performance and laboratory/flight testing. The abstracts portrayed aircraft S&T as far more performance and testing oriented, with only modest emphasis on longevity and maintenance. There are many reasons for such differences, since keywords and abstracts are applied at different levels of specificity and are used for very different purposes.

Full-text phrase proximity analysis is in its infancy but has the potential to produce extremely significant relationship results. One application whose surface has barely been scratched is the accession of complementary literatures to produce innovation and discovery from disparate disciplines.[10]

Close involvement of the technical domain expert is required in all phases of the data mining study, especially those phases in which computational linguistics plays a major role (information retrieval using iterative query, phrase frequency and proximity analyses, final integration and interpretation). From a long-range customer/user strategic viewpoint, the key output of the data mining study is not the records, computer printouts, reports, or papers. The key output is the educated expert, i.e., the technical domain expert who has had his/her horizons broadened substantially as a result of the close involvement with all stages of the data mining process and is now available to use this information and broader perspective to support management/performance of the technical area. While tools used for the data mining process are certainly important, the value of specific tools is far less than that of the expert.

However, there appears to be a steep learning curve associated with integrating the expert with the tools. The experience and training of most technical experts are focused on the technical domain, and data mining with sophisticated tools has not been part of their repertoire. Consequently, a significant amount of time is required in order for the technical experts to understand how to apply and interpret the tools. The combination of the steep learning curve and the strategic value of the educated expert suggests that the cost-effective operational mode is to retain the expert for long-term involvement with the program/topical area.

There are many potential sources of experts, and the selection of expert(s) depends on the objectives and complexity of the data mining study. For an S&T sponsor, if its Program Officers serve as the technical domain experts, the long-range involvement issue is resolved. For many reasons, including the benefits of hands-on involvement and the absence of intermediaries between the data and the user, it is preferable for the Program Officers or other users to serve as the technical domain experts if at all possible. If the data mining study is complex and projected to require substantial time from the technical expert, then a third-part intermediary, such as an external contractor, would be required. In this case, having the contractor on long-term part-time retainer to produce periodic updates and provide real-time consultation to the customer would be one mechanism to exploit the long-term involvement.

For an S&T sponsor, data mining cannot be used sporadically to realize its full benefits, but must become an integral part of the S&T sponsor's business operations. A strategic plan that presents the sponsor's textual data mining in this larger context is required to ensure that data mining integration is implemented in a cost-effective manner. Such a plan would identify the different ways data mining would support the sponsor's operations, such as planning, reviews, assessments, metrics, oversight response, etc. Each of these applications has different objectives, metrics to address those specific objectives, data requirements for each metric, different types of experts required, and different suites of data mining tools required. A strategic plan allows a top-

down driven approach to data mining, in which the desired objectives are the starting point, and the data required to satisfy the objectives can be identified, and planned for, in advance. Without such a plan, the organization is constrained by whatever data exists and has been gathered for other purposes. This bottom-up approach forces the organization to use whatever metrics the existing data will support, whether these metrics are most appropriate to satisfying the overall objectives of the application.

## 6. CONCLUSIONS

The frequency distributions of the SCI keywords associated with each paper were analyzed. An overall picture of fullerene S&T emerges, based on the journal publication results. The field may be divided into three categories, listed in descending levels of effort: materials, experiments, and theory/computation.

Overall, the field appears exclusively focused on basic research, with the majority emphasis on experiments. There are essentially no technology terms, no application terms, and very few terms relating to computation and modeling. The major focus is on C60 and C70, from metals, isomers, and polymers, although some of the higher fullerenes are alluded to as well. Some alternative sources are described, such as soot and flames. There is no indication that elements other than carbon-based are being examined for ordered fullerene-like structures.

The major geometries examined are thin-walled variants (tubules, thin films, and shells), and the major phases/structures revolve around clusters, single crystals, and gas and solid phases. A substantial amount of the research appears centered around the molecular, atomic, and electronic charge structures. Lattice/cage structures are examined heavily, as are electronic states and their associated resonances and vibrational frequencies.

While the majority of experimental activity centers on molecular and electronic spectra, there appears to be some effort on macrochemistry approaches. The probing techniques emphasize diffraction and scattering of a broad radiation spectrum. The phenomena studied focus on growth, phase transition and separation, emission over a wide frequency spectrum, and alignment and ordering of the structure at the molecular level.

High-frequency single, double, and triple word phrases from the text of the SCI database whose technical contents were deemed by topical experts to be significant were identified as the pervasive themes. From a global perspective, the SCI fullerene database has three components: two are major, and one is minor. The major categories are materials and experiments, and the minor category is theoretical/computational.

The main focus of the research is clearly experimental, with a much smaller effort devoted to theory and computational approaches.

There does not appear to be a diversity of computational approaches; the focus is on molecular orbital calculations. The experimental focus is on diverse aspects of spectroscopy and microscopy at the atomic/molecular level. Main areas of study are the diverse phase and state transitions due to the effects of incoming radiation, especially in the electron state changes. Mass and radiation spectra are the major types of outputs studied.

Experiments at the macrolevel, such as strength or fatigue tests, or chemical thermodynamics tests or techniques, are not listed.

Main material types studied are the metal compounds, with prime focus on C60 and C70, but with some attention paid to the higher fullerenes and fullerene derivatives. Macrostates studied most extensively are clusters and single crystals, mainly in thin-film and nanotube geometries. The macrolevel properties studied center around electronic primarily and optical secondarily. There is essentially negligible effort on studying mechanical and chemical/thermodynamic properties. Microstates studied center around electronic energy and transition states, at the atomic/molecular/ionic level. This couples well with the experimental focus on spectroscopy and microscopy.

Noticeable by their absence are any mention of technology or application issues. After 15 years since fullerenes were discovered, and substantial promises about new applications were made, the field remains locked in basic research at the microstructure level. Very limited geometrical configurations are mentioned, mainly focused on thin structures, and there appear to be no forays to explore similar phenomena in other materials besides carbon.

Phrase proximity analysis of the block and abstract databases for CARBON NANOTUBES yielded the following conclusions: The journal and departmental emphases are on physics, chemistry, and materials. The major performing institutions are located throughout the world. In modest contrast to the total database, Japan appears to have greater efforts than the USA in this area. In terms of sheer numbers, the Asian countries are matching the western democracies in output. Geometries focus on thin-walled tubes, experiments focus on growth and nanoscale microscopy, and computational efforts appear minimal. A large number of fullerene researchers appear fully focused on carbon nanotubes, although only one journal (*Physical Review B*) appeared to have such a fullerene subarea focus.

The main material classes examined are metals and composites, and the main carbon class examined is graphite. An application focus appears to be catalysis-related. The experimental focus is on nanoscale microscopy and diffraction, with some electrochemical discharge-oriented research as well. The major experimental purpose appears to be studying growth and deposition and some emphasis on decomposition.

This paper has presented a number of advantages of using DT and bibliometrics for deriving technical intelligence from the published literature. Large amounts of data can be accessed and analyzed, well beyond what a finite group of expert panels could analyze in a reasonable time period. Preconceived biases tend to be minimized in generating road maps. Compared to standard co-word analysis, DT uses full text, not index words, and can make maximum use of the rich semantic relationships among the words. It also has the potential of identifying low occurrence frequency but highly theme related phrases that are “needles-in-a-haystack”, a capability unavailable to any of the other cooccurrence methods.

Combined with bibliometric analyses, DT identifies not only the technical themes and their relationships but relation-

ships among technical themes and authors, journals, institutions, and countries. Unlike other road map development processes, DT generates the road map in a "bottom-up" approach. Unlike other taxonomy development processes, DT can generate many different types of taxonomies (because it uses full text, not key words) in a bottom-up process, not the typical arbitrary top-down taxonomy specification process. Compared to cocitation analysis, DT can use any type of text, not only published literature, and it is a more direct approach to identifying themes and their relationships.

The maximum potential of the DT and bibliometrics combination can be achieved when these two approaches are combined with expert analysis of selected portions of the database. If a manager, for example, wants to identify high-quality research thrusts as well as science and technology gaps in specific technical areas, then an initial DT and bibliometrics analysis will provide a contextual view of work in the larger technical area, i.e., a strategic road map. With this strategic map in hand, the manager can then commission detailed analysis of selected abstracts to assess the quality of work done as well as identify work that needs to be done (promising opportunities).

## REFERENCES AND NOTES


(1) Kostoff, R. N. Database Tomography for Technical Intelligence. *Compet. Intell. Rev.* **1993**, *4*, 1.

(2) Kostoff, R. N. Database Tomography: Origins and Applications. *Compet. Intell. Rev.* (*Special Issue Technol.*) **1994**, *5*, 1.

(3) Kostoff, R. N.; et al. System and Method for Database Tomography. U.S. Patent No. 5440481, 1995.

(4) Kostoff, R. N.; Eberhart, H. J.; Toothman, D. R. Database Tomography for Information Retrieval. *J. Inf. Sci.* **1997**, *23*, 4.

(5) Kostoff, R. N. *The Handbook of Research Impact Assessment*, 7th ed.; DTIC Report Number ADA296021; 1997. Also, available at http://www.dtic.mil/dtic/kostoff/index.html.

(6) Kostoff, R. N.; Eberhart, H. J.; Toothman, D. R. (1999a). Hypersonic and Supersonic Flow Road Maps Using Bibliometrics and Database Tomography. *J. Am. Soc. Inf. Sci.* **1999** (15 April).

(7) Kostoff, R. N.; Eberhart, H. J.; Toothman, D. R. Database Tomography for Technical Intelligence: A Road Map of the Near-Earth Space Science and Technology Literature. *Inf. Process. Manage.* **1998**, *34*, 1.

(8) Kostoff, R. N.; Eberhart, H. J.; Toothman, D. R.; Pellenbarg, R. Database Tomography for Technical Intelligence: Comparative Road maps of the Research Impact Assessment Literature and the *Journal of the American Chemical Society*. *Scientometrics* **1997**, *40*, 1.

(9) SCI. *Science Citation Index*; Institute for Scientific Information: Philadelphia, PA, 1999.

(10) Kostoff, R. N. Science and Technology Innovation. *Technovation*. **1999**, 19.

(11) EC. *Engineering Compendex*; Engineering Information, Inc.: Hoboken, NJ, 1999.

(12) Lotka, A. J. The Frequency Distribution of Scientific Productivity. *J. Wash. Acad. Sci*. **1926**, 16.

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

(14) Anwar, M. A.; Abu Bakar, A. B. Current State of Science and Technology in the Muslim World. *Scientometrics* **1997**, *40*, 1.

(15) Garfield, E. History of Citation Indexes for Chemistry—A Brief Review. *J. Chem. Inf. Comput. Sci*. **1985**, *25*, 170.

(16) Kostoff, R. N. Use and Misuse of Metrics in Research Evaluation. *Sci. Eng. Ethics* **1997**, *3*, 2.

(17) Kostoff, R. N. Citation Analysis Cross-Field Normalization: A New Paradigm. *Scientometrics* **1997**, *39*, 3.

(18) MacRoberts, M.; MacRoberts, B. Problems of Citation Analysis. *Scientometrics* **1996**, *36*, (Jul—Aug), 3.

(19) Kostoff, R. N. (1998b). The Use and Misuse of Citation Analysis in Research Evaluation. *Scientometrics* **1998**, 43 (Sep), 1.


CI990045N