Education for Information 32 (2016) 125–148
DOI 10.3233/EFI-150971
IOS Press

Bibliometrics, librarians, and bibliograms

Howard D. White
College of Computing and Informatics, Drexel University, Philadelphia, PA 19104, USA
E-mail: whitehd@drexel.edu

This paper sets forth an integrated way of introducing bibliometrics to relatively non-quantitative audiences, such as librarians and iSchool students. The integrative device is the *bibliogram*, a linguistic object consisting of a seed term and the terms that co-occur with it, ranked by their co-occurrence counts with the seed – a standard "core and scatter" distribution. While the counts and the measures derived from them are indeed central to bibliometrics, the associated terms are also important, and they exhibit distinctive features that lead to psycholinguistic insights into their distributions. This verbal side of bibliometrics is seldom highlighted when the field is introduced, yet it may be of particular interest to persons with backgrounds in the humanities. A list of words that students associate with the word "information" is presented as a reference point from psycholinguistics. Then term associations in bibliograms are illustrated with the journals that cite an author, the books assigned to a Library of Congress class, the journals that yield varying numbers of articles to three literatures defined by seed terms, the bylines that an author cites in a reading list, and the authors that are co-cited with Plato in a map of the Plato literature.

Keywords: Teachers, students, iSchools, term distributions, count distributions, ease of association

1. Introduction

It is no secret that iSchools tend to lack full courses in bibliometrics; what they offer is found in individual sessions of courses or doctoral seminars [8,40]. Perhaps the main inhibitor of bibliometric coursework has been lack of employer demand for such skills in their new hires. As a result, when iSchools do offer a course on bibliometrics (or the like), it may well be under-enrolled. Some students see it as too library-oriented; others see it as too quantitative and insufficiently linked to their job prospects. There is no obvious textbook for non-quantitative types; for instance, despite their considerable merits, Egghe and Rousseau [14] is too mathematical and DeBellis [9] is too removed from library practice. There is also the matter of faculty who can teach such a course; iSchool instructors specializing in bibliometrics are fairly rare, and rarer still are those who can popularize the material and make it relevant to students.

However, Google and the Web have diminished the need for ready-reference skills in library public services. Hence, a movement is growing to give certain academic or special librarians new skills in assembling and imparting high-quality bibliometric data [1,8,11,15,18] – the kinds of data that can support, for example, evaluation of faculty in tenure and promotion cases [4,11,37], evaluation of institutional research units [2,4,11], detection of research trends [2], and curriculum planning [32].

Other contributors to the present issue discuss aspects of this movement, building on prior writings. My intent here is to propose some ideas that might be used in introducing bibliometrics to librarians and library-bound iSchool students, who frequently come from the humanities. At present, I suspect, they are simply told how standard bibliometric measures are computed and where to find them [17]. I aim instead to set the quantitative data of bibliometrics in an unconventional qualitative framework, so as to pique their interest and, ideally, lead them to a better understanding of the subject matter. I am suggesting content and examples for lectures.

2. Two varieties of bibliometrics

Bibliometrics comes in two varieties: content-neutral and content-laden. Both derive from the fundamental linguistic object of bibliometrics, which I call a *bibliogram* [34]. Bibliograms display terms from the fields of records in bibliographic databases – the noun phrases or proper names contributed by authors, citers, editors, indexers, and librarians. The terms include such familiar data as authors' bylines, titles of published works, names of journals and publishers, names of cited authors and works, descriptors, Library of Congress and Dewey class numbers, and Library of Congress subject headings. Over time, these terms vary in their frequency of use and so can be rank-ordered, high to low. Bibliograms are thus ranked frequency distributions with (1) a verbal side – words or names – and (2) a numeric side – counts of how frequently those words or names occur in a given context.

Importantly, this context is set by a seed term that a user enters into the database – a term that defines the user's interest. The seed is also a query in the information retrieval sense. However, in bibliometrics it is not documents that are wanted (at least not initially); it is *a list of terms that co-occur with the seed.* The database responds by producing such a list and their co-occurrence counts. The ranked counts position individual terms in the overall distribution so that they can be compared and evaluated for relative importance. Fields of the bibliographic record determine the kinds of seed terms and of co-occurring terms the user can select. (For example, the seed might be an author's name, and the co-occurring terms might be the descriptors that have been assigned to that author's publications, ranked by frequency of assignment [28]). Terms may also be drawn from digitized body-text in, e.g., books or groups of related articles. In pre-computer days, the equivalents of bibliograms were manually compiled by bibliometric pioneers such as A.J. Lotka, S.C. Bradford, and George Zipf [29].

In brief, bibliograms are linguistic constructs that comprise:

- A user-supplied seed term (or group of seed terms) that sets the user's context of interest.
- Terms that co-occur with the seed across some group of records, often large.
- Counts (frequencies) of the co-occurrences by which the terms can be ranked.

The numeric side of bibliograms is the basis for content-neutral analyses – neutral in the sense that they are meaningful no matter what terms accompany them. (Some mathematical analyses rise above terms altogether – e.g., [13].) It is the numeric side that yields the measures now most commonly associated with bibliometrics, such as the h-index or the journal impact factor.

The verbal side of bibliograms appears in content-laden analyses – for instance, maps of research specialties – in which interpretations of related terms are the main goal. The verbal side may reflect the intellectual careers of authors and, more broadly, the historical development of literatures in specific fields of science and scholarship.

The word "bibliogram" was coined to stress the *terms* in the distributions equally with the *counts*. Many bibliometricians, seeking generalizable results, focus on the counts alone. When plotted, the count distributions in bibliograms have a distinctive shape, which mathematical bibliometricians characterize with content-neutral phrases such as "empirical hyperbolic," "high skew," "rank-frequency," "scale-free," "Pareto," "reverse-J," or "power law." Here, the focus will instead be on the content-laden term structures that the counts create.

3. Associating terms

Bibliograms are fittingly described as term association structures. They share features with the word association structures studied in psycholinguistics [34,35]. Usually they represent the combined associations of many persons, but they can also result from the associative behavior of individual authors over time. They thus invite *psychological* explanations that are absent in content-neutral bibliometrics, which explains count distributions probabilistically.

The extracts from bibliograms in the present paper make strong and weak term associations evident, which is a relatively concrete way of teaching bibliometrics. What is being associated at various strengths are the seed term and each co-occurring term. In psycholinguistics, people are given a stimulus word and told to respond with any other word that occurs to them. In bibliometrics, the seed term resembles the stimulus word, and the co-occurring terms resemble the response words. Both psycholinguistic and bibliometric distributions moreover exhibit "core and scatter" structure.

To illustrate this structure, Table 1 displays a word association list from the Edinburgh Associative Thesaurus [12]. It shows the 44 words that 84 Scottish students gave in response to the stimulus word "information." The rank-ordered response frequencies in the count column are very unevenly distributed and can be placed in three core-and-scatter zones. Dividing 100% of all 84 responses roughly into thirds, the frequencies of the four most mentioned words ("knowledge" through "data") sum to 29 or about 34% of all responses. The frequencies for the next eight words sum to 24 or 29%, while the remaining words, with one response each, sum to 31 or

Table 1
EAT responses to the stimulus word "Information"

Count: Responses	Terms: Associated words
9	Knowledge
7	Desk
7	News
6	Data
5	Office
4	Bureau
3	Please
2	About
2	Centre
2	Facts
2	Given
2	Notes
2	Service
1	Ask, Bits, Blank, Carrier, Communication, Computer, Documents, Elusive, Enquiry, Explanation, Form, Formula, Give, Guide, Helpful, Information, Informer, Interest, Know, Known, Leading, Lecture, Letter, Necessary, No, None, Now, Only, Police, Prisoner, Processing

37%. The four words with the most mentions can be considered the *core* of the structure, while the remaining mentions are *scattered* over increasingly large numbers of words – eight and 31 respectively.

Figure 1, which graphs the response counts in Table 1, illustrates core-and-scatter structure in a small dataset. Table 1's core words are represented by the four highest points at left, and the 31 "onesies" run rightward along the bottom. The graph's "reverse-J" structure is related to the power law curves that are a routine finding in mathematical bibliometrics. This structure turns up, for example, in studies of author productivity, author citedness, book sales, collection sizes, literature sizes, and customer use of libraries. I have plotted a power-law curve through the data points in Fig. 1 with DeltaGraph software.

To account for the structure seen in Table 1, I will introduce a psychological claim: the more frequently a word is mentioned, the easier that word is to associate with the stimulus word. In some association studies responses are timed, and those made faster, on average, are presumably easier to give than slower ones. Here, response frequencies substitute for time measures. The student, I trust, will be able to look at the ranked frequencies and interpret higher-ranked associations more readily than lower-ranked ones.

Interpretation is based both on knowing what words mean and on knowing facts about the world. In many academic fields, for example, the word "information" is constantly linked to "knowledge" and "data," and is discussed with them as a trinity. Real life also provides ample experience with "information desks" and information as "news." Similar explanations can be quickly made up for most of the other words appearing more than once ("office," "bureau," and "centre" are more Scottish than

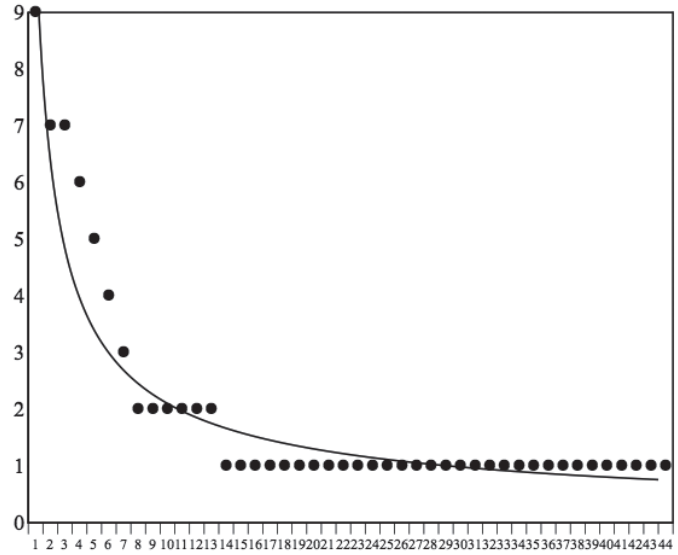

Fig. 1. Plot of EAT responses to the stimulus word "information".

American in flavor). However, as the mentions decrease, the words become increasingly idiosyncratic. Some that were mentioned only once are somewhat puzzling. Compare, say, "necessary, no, none, now" with "knowledge, news, desk, data." It is not that the first group seem *totally* unlikely, merely much *less* likely than the second. We can detect differences in the plausibility of the word associations, which argues that we can judge them on a mental ordinal scale that runs from *very interpretable* to *not interpretable*. A very interpretable association is one that can be confidently explained on sight. Even better is an explanation that is not too idiosyncratic – one that quickly makes sense to others.

We can also judge terms in bibliograms as the counts descend from high to low. A basic idea in teaching bibliograms is that the count that accompanies each term *weights* that term. (Term weights are a standard technical feature in information retrieval, but the notion is useful in bibliometrics as well.) The higher the weight, the stronger the association with the seed. Ranking terms is equivalent to ordering them by their strength of association. A term highly weighted with respect to the seed tends to be easily interpretable with respect to it on our internal ordinal scale. Moreover, *groups* of highly weighted terms tend to concentrate meanings with respect to the seed in a way that lower-weighted terms do not, as will be shown. Terms with low weights, in contrast, tend to be obscure with respect to the seed. It is harder to explain their relation to it, although the relation undeniably exists. The following pages give examples for use by teachers.

Table 2
Core and scatter terms in WoS journals citing works by John M. Swales as of mid-2015

Counts: Citations	Terms: Journal titles
243	English for Specific Purposes
87	Journal of Pragmatics
84	Journal of English for Academic Purposes
74	TESOL Quarterly
73	Written Communication
71	Applied Linguistics
69	Journal of Second Language Writing
61	Iberica
39	Journal of Business and Technical Communication
36	Discourse Studies
31	Revista Signos
28	Discourse Society
27	Research in the Teaching of English
24	Modern Language Journal
23	Text Talk
20	System
1	Psychosomatic Medicine
1	Public Historian
1	Quality in Health Care
1	Quarterly Journal of Speech
1	Reading Teacher
1	Regional Studies
1	Research Evaluation
1	Research Policy
1	Respiratory Medicine
1	Rethinking History

4. Example: Journals citing an author

The seed of the first bibliogram displayed here is an author's name, John M. Swales. A professor of English, he is well known for his publications on the rhetoric and genres of academic discourse (e.g., [26,27]). His work also bears on teaching students, including those from non-Anglophone countries, how to compose research papers and to write English in a scholarly or scientific style.

Table 2 provisionally answers the question, "In what journals is Swales most cited?" In July 2015 I put this question to the Social Sciences Citation Index and Arts & Humanities Citation Index as combined in the Web of Science. Since WoS requires seed names to be shortened to surname and initials, I entered "Swales J OR Swales JM" – he is cited both ways – as a cited author in a Cited References search. This retrieved a large set of articles, many of them clearly involving the wrong Swales. "Swales's J" brought in numerous journals citing "Swales JD" and "Swales JK," and so I eliminated those authors from the set with NOT logic. After various trials, I decided not to use WoS check boxes to limit the set further by subject, because I know Swales's work to be of cross-disciplinary interest. I then used the Analyze Results module on the final set – specifically the "Rank records by source titles" command –

to produce the complete bibliogram, down to the many journals that each cite Swales once. Table 2 displays the journals with counts of at least 20 – some or all of which would be in the core of the bibliogram.

The final set gave a simple citation count for Swales, but the extract in Table 2 (which could be expanded) is more informative. First, if requested by Swales or some other party, it would have human interest as news. Second, it coherently expresses Swales's intellectual world in terms of journal titles. His work very plausibly has its greatest impact in applied linguistics and communications journals – an example of top-ranked terms being jointly quite intelligible.[1] Third, the journals suggest that citation credits are going to the right author and not to someone with the same name and initials.

To highlight the difference between top-ranked and bottom-ranked terms in the Swales bibliogram, Table 2 also has 10 journals arbitrarily chosen from the 58 that cited him once. (Similar high-low contrasts appear in some of my other papers [29,35,36].) It will be seen that the disciplinary coherence of the top-ranked journals has disappeared. Since these titles come from several disciplines, they are more miscellaneous than the top-ranked titles, as usually happens with extreme scatter terms. Recall the odd mix of words that received one mention each in Table 1.

It remains to say that the bibliogram underlying Table 2 has not been checked for errors. Abbreviations like "Swales JM" and especially "Swales J" are not guaranteed to refer solely to the author intended. Works by one or more different authors with the same name-string (homonyms) might be inflating the counts and bringing in erroneous journals. This could be important, for instance, if one wants to use the WoS Analyze Results module to reveal the *breadth* of Swales's impact by making bibliograms of the Research Areas or the even broader WoS Categories to which the journals that cite him are assigned. Including the wrong Swales would lead to erroneous bibliograms on these higher subject levels. Therefore, detailed checking is needed to make sure that citation credits are correct.

Attention to bibliographic detail is a stock-in-trade among librarians. If they are going to assist in practical bibliometric analyses, one of their main concerns should be errors in the data that affect the accuracy of the counts (see, e.g., [25] and "Homonyms and Allonyms" in [30]). Such errors are known to exist in all the databases mentioned here. They particularly affect proper nouns such as authors' names and the names of books, serials, and organizations.

Perhaps more important in Swales's case are possible sins of omission. His work is relevant to researchers well beyond the fields of linguistics and communications.

[1] Domain experts could interpret the non-transparent names, but to clarify them for others, *TESOL* stands for "teaching English to speakers of other languages." *Iberica* is the journal of the European Association of Languages for Specific Purposes. *Revista Signa* publishes articles in English and Spanish on linguistics, psycholinguistics, text linguistics, discourse linguistics, and applied linguistics. The international journal *System* specializes in applications of educational technology and applied linguistics to problems of foreign language teaching and learning.

Table 3
Books in LC class Z711 (Reference Librarianship) most purchased by OCLC libraries in the 1980s

Counts: Holding libraries	Terms: Book titles
546	*Learning the library*; Concepts and methods for effective *bibliographic instruction*
544	*Bibliographic instruction*; A handbook
479	Planning the *library instruction* program
455	Reference and online services handbook: Guidelines, policies, and procedures
445	Evaluating *bibliographic instruction*; A handbook
445	Finding the source: A thesaurus index to the reference collection
418	Theories of *bibliographic education*; Designs for teaching
387	Reference services and *library instruction*
367	*Teaching library use*; A guide for *library instruction*
364	*Bibliographic instruction*; The second generation

Although many of these other fields and specialties are covered by the two databases I chose to search, my initial search strategy eliminated the Science Citation Index altogether, which may have eliminated some science journals that repeatedly cite him. It would be wise to check the SCI for additional citations to his work that the present analysis missed. (For example, he is cited in medicine – not as a medical researcher but as an authority on the structure of research articles.) There is also the glaring omission of the many *books* that cite him; these can only be found in databases outside the ones I searched, notably Google Scholar. Librarians typically know things like this.

5. Examples: Libraries holding books

Currently, librarians tend to view bibliometrics as almost exclusively a matter of citation counts (see, e.g., [17]). Such counts are indeed a very important part of the field, but they should be understood as a manifestation of a more general process. My next examples of bibliograms for librarians and iSchool students are therefore taken from OCLC's WorldCat, a database not usually associated with bibliometrics. Although much of the data on books and serials in WorldCat originates with authors and publishers, it is librarians who index and classify those titles by subject. Thousands of other librarians use WorldCat in contexts such as reference work, copy cataloging, and interlibrary loan. This may heighten its appeal as a source of bibliograms.

The great majority of items in WorldCat are books. Table 3 displays the 10 books in Library of Congress class Z711 that were most frequently held by OCLC member libraries during the 1980s. (For example, 544 libraries reported cataloging and holding *Bibliographic Instruction: A Handbook*.) The top 10 are, once again, a tiny part of the full bibliogram, which has more than 2,000 book titles.

The seed term was Z711, which stands for Reference Librarianship as an academic subject. The book titles are the terms that co-occur with it. The data were easy to

gather because WorldCat by default ranks the records of books co-occurring with the seed by their holdings counts, just as they appear in the table. I have simply reduced their full bibliographic entries to counts and titles. The counts are not current, but that does not matter here.

Table 3 is a remarkable example of how groups of terms concentrate meaning with respect to the seed in a way no single term could do. The italicized phrases phrases in top-ranked titles do not mechanically *match* the seed term, they jointly *thematize* the dominant concern of reference librarianship (Z711) in the 1980s. Clearly, the preoccupation back then was teaching people how to use libraries, also known as bibliographic instruction. Yet no one planned this result; library holdings counts are wholly blind to content. Rather, the extract makes visible an editorial process in which multitudes of decisions by collection development librarians (and earlier by authors, publishers, LC classifiers, and book wholesalers) converge as if in response to a zeitgeist. Books with certain title terms are elevated so that they reveal how the seed is construed during a particular time. As one goes lower in the ranking, the thematization seen in the top ranks becomes more diffuse or disappears altogether.

The larger point is that bibliograms in general concentrate meaning at the top, regardless of the seed term chosen or the co-occurring terms requested. That is, whatever the bibliogram, the topmost co-occurring terms tend to be interpretable in light of the seed term, at least by those with the requisite domain knowledge. This includes bibliograms based on citation counts, as seen with the Swales example in Table 2.

It may seem strange that *high citation counts* and *high library holdings counts* have similar effects at the tops of bibliograms, but they do. In fact, citation counts and holdings counts have enough in common that I have called the latter *libcitations* (librarians' citations) to bring out the parallel [37]. Citations and libcitations both make "intellectual incorporations" explicit: citations show existing texts being incorporated in new writings for potential readerships, while libcitations show existing texts being incorporated in new library environments for potential customers.

One might wonder, however, whether the effect in Table 3 is merely a fluke. Table 4 accordingly presents the top 10 titles for the same seed, Z711, as of a few years ago. (At this writing, the top titles remain largely those seen in Table 4.) All the leading books in Table 3 are still held by hundreds of libraries, but they have disappeared from the top 10, since bibliographic instruction is by now routine. Instead, a new theme is italicized: reference service quality and customer relations, especially the handling of difficult customers and problem situations. (Five of the 10 books deal in some way with this ominous topic.) Times and concerns have changed, but the thematizing effect re-emerges.

The thematic information in Tables 3 and 4 is obviously very shallow. But that is to be expected; dealing as it does with entire literatures, librarianship is necessarily given to statements at a high synoptic level – to indexes, catalogs, summaries, overviews. Moreover, these statements always presume that users can drill down to finer levels of detail as needed, until they actually *read body text*. By design, shallow information leads to deeper information.

Table 4
Books in LC class Z711 (Reference Librarianship) most purchased by OCLC libraries in the 2000s

Counts: Holding libraries	Terms: Book titles
1419	*Patron behavior* in libraries: A handbook of *positive approaches* to *negative situations*
1406	Dealing with *difficult people* in the library
1325	*Customer service excellence*: A concise guide for librarians
1316	Assessing *service quality*: *Satisfying the expectations* of *library customers*
1105	The American Library Association guide to information access: A complete handbook and directory
1090	Managing the reference collection
1017	Serving the *difficult customer*: A how-to-do-it manual for library staff
965	Library research models: A guide to classification, cataloging, and computers
938	Defusing the *angry patron*: A how-to-do-it manual for librarians and paraprofessionals
933	It comes with the territory: handling *problem situations* in libraries

6. Example: Core, scatter, and Bradford

Core and scatter have their conceptual roots in a bibliometric classic – S.C. Bradford's 1934 study of the pattern in which scientific journals contribute articles to the literature of a subject [5]. The two subjects he examined were Lubrication and Applied Geophysics as research fields. For example, using "lubrication" as a seed (he did not call it that), he created a bibliography of journal articles indexed by or related to that term. Rank-ordering the journals by the number of lubrication articles they published – their yield or productivity – he found that approximately a third of the articles came from a small core of journals, another approximate third were scattered across a greater number of journals, and the final third were scattered across far more journals still.

Having thus divided his journals into a core zone and two scatter zones, he concerned himself with finding a mathematical relation between the number of journals in each zone – the zones of what became known as a Bradford distribution. The relation he published ("Bradford's law of scattering") has been much discussed and modified [7,10,21,23]; while not a law, it is a rough empirical regularity that has now been folded into more sophisticated work on power-law phenomena.

Because bibliometric distributions other than Bradford's can be modeled with the same math, contemporary bibliometricians such as Egghe and Rousseau [14] treat the idea of "journals producing articles" as simply one example of "sources producing items." Both sources and items can be counted, and sources can be ranked by the number of items they yield. Thus a more inclusive formulation of Bradford's core-and-scatter idea is:

– *Core* consists of the relatively *few* sources that yield *many* items and so are top-ranked.

– *Scatter* consists of the relatively *many* sources that yield *fewer and fewer* items and so are progressively lower ranked, down to a long tail of sources that yield only one item each.
– It takes *many* sources in the scatter zones to produce items as numerous as those produced by relatively *few* sources in the core zone. The sources in core can be said to yield "disproportionately" many items.

Emphasizing yield, however, keeps the focus on the numeric side of bibliograms – the ranked counts. Shifting attention to the ranked terms on the verbal side, one can say that core concentrates their *meaning*, while scatter diffuses it. This notion was explored at length, with examples, in a review by White and McCain [38]. Two brief quotes: "... a term of one kind can be converted into a core and scatter of associated terms (of the same or different kinds) with high interpretability" (p. 125). "Furthermore, *ranking by frequency of co-occurrence brings out additional meaning*, meaning that is obscured if the associated terms are simply alphabetized" (pp. 127–128, italics in original). Various ways of measuring the size of the core and scatter zones have been proposed, but, as a complement to these, one can also check the interpretability of terms in the zones. Thus, recasting Bradford, one can say:

– A subject term used as a seed will tend to concentrate obviously related terms from journal titles in a core zone, but less so in scatter zones, where journal titles harder to relate to the seed proliferate.

As we saw with the WorldCat examples, the zone-forming process involves the choices of many persons over time – in this case authors who decide to submit their papers to certain journals, and journal editors who decide to accept or reject the authors' submissions. Everyone is motivated by perceptions of verbal appropriateness. In other words:

– Concentrated journal terms tend to be relatively *easy to associate with* the seed term, which is why the journals yield disproportionately many articles to its bibliography.

Bradford, for his part, reported only the numeric side of his analyses. He might have published at least the core journals in Lubrication and Applied Geophysics, but apparently he considered only the math of the zones to be of interest (bibliometricians seldom think like journalists), and his title data seem to have been lost. Therefore we cannot examine the verbal side of what were, in effect, early bibliograms.

Out of curiosity, however, I once did a search on Dialog that used the seed term "lubrication" as it occurred in article titles in the Science Citation Index. From the set of articles thus formed, I ranked by yield the journals that published them. (See [33] for a map and a fuller account of the lubrication data.) Table 5 contrasts the top-ranked journals with some in the middle ranks. It is immediately plain that the top-ranked journals not only have richer yields, their titles also tend to be more transparently associated with the seed term.

Table 5
Core and scatter terms in journal titles retrieved in Science Citation Index with seed term "lubrication"

Counts: Articles	Terms: Journal titles
236	Journal of *Tribology*
192	*Wear*
167	*Tribology* Transactions
138	Journal of Japanese Society of *Tribologists*
135	*Lubrication* Engineering
102	*Tribology* International
102	Proceedings of the Institution of Mechanical Engineering
40	*Tribology* Letters
30	Abstracts of Papers of the American Chemical Society
25	Journal of Materials Processing Technology
21	Industrial *Lubrication* and *Tribology*
7	Die Casting Engineer
7	Langmuir
7	Journal of Applied Polymer Science
6	Journal of Non-Newtonian Fluid Mechanics
6	International Journal of Engineering Science
6	CIRP Annals – Manufacturing Technology
5	Journal of Rheology
5	TAPPI Journal
5	ZKG International
5	Chemistry and Technology of Fuels and Oils
5	British Journal of Anaesthesia
5	Journal of Dental Research

When Bradford [5] discussed scatter in the literature of a subject, he did not define "subject." Hjørland and Nicolaisen [16] have sought clarification by proposing and glossing three different kinds of subject scatter in Bradford distributions. They also propose that subject scatter be operationalized by looking for verbal patterns in citations. The journal titles in Table 5 are from citations, and to a degree they illustrate the three kinds of scatter, as bulleted:

- *Lexical*, glossed as the scattering of *words* in texts. The seed term "lubrication" formed a set by retrieving instances of itself in article titles. As it happened, this also brought in two instances of it in the top 11 journal titles. (Linguists would call the word "lubrication" a *type*, and the various instances of it in the retrieval are *tokens* of that type.)
- *Semantic*, glossed as the scattering of *concepts* in texts. Hjørland and Nicolaisen give the example of nouns and their synonyms that appear in the classification or indexing schemes of disciplinary domains. In Table 5 "Tribology" and "Tribologists" are domain-defining terms superordinate to "lubrication." Tribology, according to the Apple Dictionary, is "the study of friction, wear, lubrication, and the design of bearings; the science of interacting surfaces in relative motion." Other titles relate to "lubrication" at even higher superordinate levels, such as "Mechanical Engineering" and "Materials Processing Technology."
- *Subject*, glossed as the scattering of *items useful to a given task or problem*. The

Table 6
Core and scatter terms in journal titles retrieved in the SSCI and AHCI databases with seed term "Kierkegaard"

Count: Articles	Terms: Journal titles
94	International Journal for *Philosophy* of *Religion*
55	*Religious* Studies
47	Tijdschrift voor *Filosofie*
41	Heythrop Journal – A Quarterly Review of *Philosophy*
38	International *Philosophical* Quarterly
36	Review of *Metaphysics*
33	*Philosophy* Today
28	Revue *Philosophique* de la France et de l'Étranger
26	TLS – The Times Literary Supplement
25	Revue des Sciences *Philosophiques* et *Theologique*
24	Scandinavian Studies
1	Southern Humanities Review
1	Southern Review-Adelaide
1	Studia Phaenomenologica
1	Studies in Romanticism
1	Symposium – A Quarterly Journal in Modern Literatures
1	Synthese
1	Tempo
1	Teorema
1	Textual Practice
1	Thalia – Studies in Literary Humor

top-ranked titles in Table 5 are here more indirect, but *Wear* names a problem for which lubrication is a useful solution.

The mid-ranked titles in Table 5, when not cryptic to the outsider, are not at all specific to "lubrication" and come from a variety of domains. They confirm that, at some point in a bibliogram, the concentration of meaning seen in core will give way to the greater diffuseness seen in scatter.

7. Examples: More evidence from titles

The next two examples show that effects such as those seen in Table 5 are not confined to literatures from science and technology; they occur in humanities literatures as well. Tables 6 and 7 were both created about five years ago in Arts & Humanities Citation Index on Dialog. They further illustrate the concentration and diffusion of meaning in core-and-scatter zones.

In the WorldCat examples given in Tables 3 and 4, the associations were between the seed term Z711, Reference Librarianship, and book titles that were *subordinate* to it, in the sense of thematizing specific aspects of it. In contrast, the journal titles seen in Tables 6 and 7 are *superordinate* to their seed terms, indicating the disciplines from which the seed terms come, just as the journals in Lubrication did. (It remains to be seen whether such hierarchical effects regularly occur across literatures.)

Table 7
Core and scatter terms in journal titles retrieved in the SSCI and AHCI databases with seed term "Fairy Tales"

Counts: Articles	Terms: Journal titles
285	*Fabula*
57	Zeitschrift für *Volkskunde*
48	TLS – The Times Literary Supplement
37	Journal of American *Folklore*
34	*Osterreichische* Zeitschrift für *Volkskunde*
28	New York Times Book Review
27	*Children's Literature* In Education
25	*German* Quarterly
25	*Volkskunde*
24	*Folklore*
21	Western *Folklore* s
21	Zeitschrift für *Germanistik*
2	Critique – Studies In Contemporary Fiction
2	Cuadernos Hispanoamericanos
2	Dance Magazine
2	Dancing Times
2	Dickens Quarterly
2	Dix-Septième Siècle
2	Down Beat
2	Dreamworks
2	Euphorion – Zeitschrift für Literaturgeschichte
2	Film Quarterly

In Table 6 the seed term was simply Kierkegaard, the Danish philosopher, as a subject of study. The 10 journals with the highest yields reveal that he is indeed studied in philosophy but more particularly in the philosophy of religion, which makes sense to anyone who knows his writings. The same disciplinary identifications continue to occur in titles well below those listed. Table 6 also brings out Kierkegaard's importance in various languages and cultures; journals in English, Dutch, and French are seen in the top 12, and journals in English, French, and German appear in the next several ranks not shown.

Table 7 had as its seed the phrase "Fairy Tales," again as a subject of study. The discipline identified here is Folklore Studies, but with a Germanic cast (*Volkskunde* means "folklore" – e.g., Grimm's tales). We see journals both *in* German and *about* German language and literature (*Germanistik*). The top-ranked *Fabula* is the multilingual journal of the International Society for Folk Narrative Research.

The titles of journals contributing one article apiece to the Kierkegaard literature suggest his importance to the humanities in general, but the cohesiveness of the core journals is lost. The same loss appears in the journals that contribute two articles each to the Fairy Tales literature.

8. Example: Citations yielding bylines

In Table 8, I use the core-and-scatter framework to examine some rather unusual

Table 8
Bibliogram of bylines retrieved from readings is Deirdre Wilson's course "Pragmatic Theory"

Count with seed	Cumulative count by zone	Terms: author bylines	Cumulative % of bylines	Cumulative % of counts
16	16	Sperber, D. & Wilson, D.	3	20
9	*25*	Wilson, D. & Sperber, D.	6	31
7	7	Carston, R.	8	40
6	13	Grice, H. P.	11	47
5	18	Wilson, D.	14	53
3	21	Sperber, D.	17	57
3	24	Recanati, F.	19	60
3	27	Levinson, S.	22	64
2	*29*	Wilson, D. & Carston, R.	25	67
1	1	Asher, N. & Lascarides, A.	28	68
1	2	Bach, K.	31	69
1	3	Bach, K. & Harnish, R.M.	33	70
1	4	Baron-Cohen, S.	36	72
1	5	Beveridge, M. & Marsh, L.	39	73
1	6	Campbell, R. & Bowe, T.	42	74
1	7	Clark, E. & Clark, H.	44	75
1	8	Clark, H. & Gerrig, R.	47	77
1	9	Gernsbacher, M. (ed.)	50	78
1	10	Gibbs, R.	53	79
1	11	Glucksberg, S.	56	80
1	12	Happé, F.	58	81
1	13	Katz, J.	61	83
1	14	Keenan, E.O.	64	84
1	15	Kreuz, R. & Glucksberg, S.	67	85
1	16	Kumon-Nakamura, S., et al.	69	86
1	17	Langdon, R., et al.	72	88
1	18	Lascarides, A., et al.	75	89
1	19	Martin, R.	78	90
1	20	McCawley, J.	81	91
1	21	Morgan, J. & Green, G.	83	93
1	22	Neale, S.	86	94
1	23	Searle, J.	89	95
1	24	Swinney, D.	92	96
1	25	von Frisch, K.	94	98
1	26	Wharton, T.	97	99
1	*27*	Wilson, D. & Matsui, T.	100%	100%
$N = 81$	*3 Zones = 81*	*36 Bylines*	$N = 36$	$N = 81$

data: *citations* as sources that yield *bylines* as countable items [35]. The citations come not from multiple persons but from one author, Deirdre Wilson. In the field of linguistic pragmatics she is famous as the co-creator, with Dan Sperber, of Relevance Theory (RT), now an international research specialty.

The seed term of the bibliogram is "Pragmatic Theory," the name of a linguistics course that Wilson taught online in 2007 for University College London, where she is now a professor emeritus [39]. The co-occurring terms are bylines – authors singly and in combination – taken from the readings she assigned or recommended at the

end of each lecture. In all, there are 81 citations distributed very unequally over 36 bylines, a few of which appear on more than one publication. Her bylines with Sperber differ in who gets first-author credit.

When individual authors are studied as citers (in, e.g., [28,29,31,33]), frequency counts for their citees are usually obtained from the big citation databases. Here, I assembled the data by hand. (Wilson's lectures were publicly available for a time on the Web.) Librarians will not create bibliograms like this, of course, but the Wilson example has features typical of many thousands of citers and gives insights into their behavior.

Lists of cited names admittedly require domain expertise to interpret. To someone versed in Relevance Theory, however, many of the names in Table 8 make instant sense. They imply *works* to which the names are attached, and the works make the names explicable. The names can also be construed as *persons*, and knowledge of authors as persons adds to one's explanatory powers [29]. However, the claims made here do not require any knowledge of RT to understand.

Taking the numeric side of Table 8 first, the bylines have been ranked by their co-occurrence frequencies and divided Bradford-style into three zones, each comprising roughly a third of total citations. The results are quite like those in Table 1:

- In descending order, two bylines yield 25 citations; seven bylines yield 29; and 27 bylines (the "onesies") yield 27. The total yields are in italics in the second column.
- The cumulative percentages at right show that only 6% of the bylines yield almost 31% of Wilson's citations, and only 25% of the bylines yield 67% of her citations. The latter result is reminiscent of the 80/20 rule or "Pareto principle," in which 80% of some outcome is produced by 20% of the causal agents. Such distributions are broadly characterized as power-law phenomena.
- Figure 2a reveals the Wilson distribution to be similar is shape to the EAT distribution in Fig. 1. A power-law curve has again been added with DeltaGraph. The point once again is to suggest relative ease of association as the force behind the two distributions.
- Figure 2b uses the same data as 2a but has the ranks on the vertical axis and the frequency of co-occurrence scale on the horizontal. Sometimes power-law researchers place their axes as in 2a, and sometimes as in 2b, depending on what they want to show. Either way, an identical power-law curve fits the data.

In choosing her readings, Wilson did not create the distribution in Table 8 intentionally; she was simply writing and citing. Her course is called "Pragmatic Theory," which licenses a great variety of possible topics and reading assignments. However, she very naturally centered the course on ideas most salient to her: those of Dan Sperber and herself as they built upon but substantially revised the ideas of H. Paul Grice, a philosopher of language particularly important in pragmatics. (These ideas involve the role of *inference* in human communication – its role in enabling hearers to derive what speakers mean from what they actually say.) The item most frequently cited in

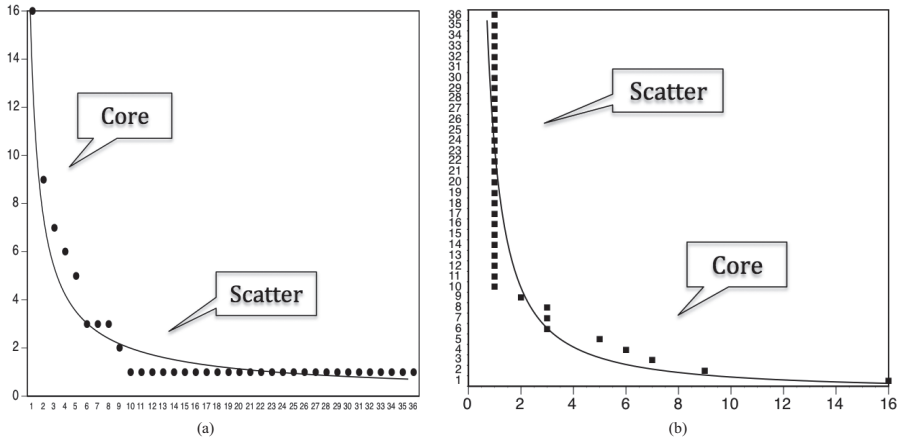

Fig. 2. (a) Plot of bylines cited in "Pragmatic Theory" course; (b) Same data as Fig. 2a but with the axes reversed.

all of Wilson's readings is *Relevance: Cognition and Communication*, the book she and Sperber first published in 1986. (He is first author.) This is the work that brought Relevance Theory to a wide readership, and Wilson's course introduces that theory, which she and Sperber have continued to develop in many publications. As such, 25 of her references (the core in the bibliogram) are to writings on RT that she and Sperber co-authored.

Besides Grice, and Sperber and herself singly, other authors that Wilson re-cites are Robyn Carston, François Recanati, and Stephen Levinson. Carston, an established relevance theorist like Wilson, is cited for her book *Thoughts and Utterances; The Pragmatics of Explicit Communication*. She and Wilson also did a paper together that Wilson cites twice. Recanati and Levinson are leading representatives of linguistic pragmatics *outside* RT; their works are useful in contexts of comparison and contrast.

The items cited once are articles and books that Wilson invoked as needed to document points or provide background. Most are from the pragmatics literature outside RT. When their titles are spelled out, they exhibit the miscellaneousness of scatter.

The writings in Table 8 are connected both explicitly, by language in common, and implicitly, by conjunctions of ideas that may be inferred. But beyond these intellectual links, there are *social* connections between Wilson and other authors as persons. Carston, for example, was Wilson's doctoral student. Social ties of various kinds among citers and citees are very common and are found especially in cores. The co-authorship relation is only the most obvious example. Authors do not cite other authors *solely* on the basis of social ties, but such ties do add to the psychological prominence of these other authors – the ease with which they spring to mind.

The strongest of ties connect an author's own works. If an author's citees are ranked by frequency of citation, self-citations are usually most frequent of all – "the core of the core." The cause is rarely egotism in a bad sense; authors are merely binding their works into greater wholes in the developing oeuvre. They are uniquely qualified to see how the latest writing grows out of earlier ones, and they want others to see it too.

All this is by way of explaining the shape of the bibliogram in Table 8. The three zones reflect the notion of building the course from the inside out, psychologically speaking. Wilson instinctively chose as readings a group of writings she already knew well. Among them, writings of her own topped the list. Other writings near the top were also easy for her to associate with her own interpretation of "Pragmatic Theory," and she cited them more than once. The principle this illustrates is economization of effort at the individual level. Wilson seeks to provide her students with relevant readings. Thus, the readings she associates with her course have, in her view, the richest implications for it while simultaneously costing her the least effort to cite. This is not laziness; it is a way of efficiently allocating attention that is built into human cognition [35].

Wilson is citing in a context outside the formal literature. But authors also tend to economize on effort when they cite in formal publications; it is a general principle of behavior. If a *citing* author is the seed term and *cited* authors are the co-occurring terms, then greater ease of association is shown by, for example:

- Disproportionate self-citation.
- Disproportionate citation of selected acquaintances.
- Disproportionate citation of certain orienting figures known only from reading – authors the citer has not or could not have met.

If a citing author is the seed term but the co-occurring terms are cited *works*, then greater ease of association is shown by, for example:

- Emergence of "personal anthologies" of disproportionately cited works – favorites from past reading that are plugged in at every opportunity.
- Repeated citation of the same works to symbolize the same concept.
- Repeated agreement in vocabulary between citing and cited works – that is, matching, near-matching, or semantically close terms in titles, abstracts, and full texts.

These points are taken from a paper that uses RT itself to explain them [35]. More broadly, they introduce aspects of citation theory with which novice audiences may be unfamiliar.

9. Example: Bibliograms and mapping

As mentioned earlier, terms from bibliograms can be mapped and their positions interpreted. An example of a co-citation map may clarify such visualizations for novices.

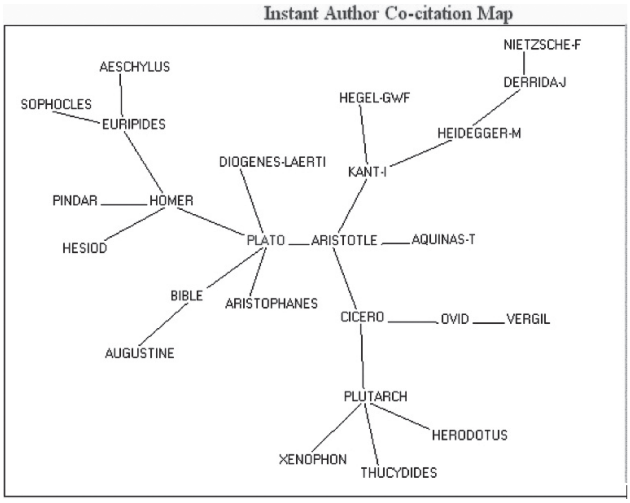

Fig. 3. Map of authors co-cited with Plato.

Figure 3 is a map of late 20th-century scholarship on Plato [6]. It is technically a Pathfinder Network (PFNET), a graphic developed by psychologists to display cognitive associations among words [24]. The map was made from a 10-year file of the Arts & Humanities Citation Index (AHCI), which records citations *from* articles in humanities journals *to* works of any sort. Figure 3 is centered on Plato, but any author with sufficient citations in a database can be mapped in the same way.

The AHCI file, comprising citing articles and cited works from 1988–1997, was given as a research grant to my college at Drexel University by Thomson Reuters. The mapping software and interface were created at Drexel under the name of AuthorWeb [19,31]. A similar system could map citation data from books as well as articles and for periods longer than 10 years; AuthorWeb has demonstrated the concept.

The map in Fig. 3 addresses the question: With what authors was Plato most studied over the decade? Or more precisely, since the names from Homer to Derrida stand for oeuvres: With whose works have Plato's works been most studied over the decade? The question is addressed by revealing patterns of author co-citation in humanities journals. Two authors are co-cited when *any work(s) by one* and *any work(s) by the other* are cited in the same article. Suppose, for instance, that an article refers to three of Plato's dialogues and a play by Euripides. This would increase the co-citation count for Plato and Euripides by one, since in AHCI it is the citing article that is counted, not the references it contains. A co-citation count for any pair of authors in AHCI is the total number of articles that cite them jointly. Co-citation counts are created by citers in general and represent their pooled associations over time, which of course vary in frequency.

	PLATO	ARISTOTLE	PLUTARCH	CICERO	HOMER	BIBLE	EURIPIDES	ARISTOPHANES	XENOPHON	HERODOTUS	AUGUSTINE	KANT-I	AESCHYLUS	THUCYDIDES	SOPHOCLES	OVID	HESIOD	DIOGENES	HEIDEGGER-M	DERRIDA-J	NIETZSCHE-F	PINDAR	HEGEL-GWF	VERGIL	AQUINA
PLATO	0	1940	872	742	664	566	532	532	499	442	444	391	385	371	374	350	346	339	317	316	306	308	279	279	269
ARISTOTLE	1940	0	864	917	531	554	409	446	473	544	423	595	317	312	374	253	271	326	413	290	229	284	383	223	744
PLUTARCH	872	864	0	1046	535	395	427	448	503	220	582	28	276	264	447	444	265	314	20	25	265	28	27	372	38
CICERO	742	917	1046	0	329	431	218	180	213	476	217	85	118	125	156	645	162	264	46	59	115	58	61	615	142
HOMER	664	531	535	329	0	233	552	333	250	105	394	24	400	416	224	446	513	121	33	65	404	60	35	568	20
BIBLE	566	554	395	431	233	0	126	83	133	908	141	173	86	90	62	257	98	81	118	118	59	126	112	266	405
EURIPIDES	532	409	427	218	552	126	0	402	258	52	314	11	475	491	224	257	277	90	13	24	307	51	12	207	5
ARISTOPHANES	532	446	448	180	333	83	402	0	307	45	297	6	270	264	263	140	193	107	4	11	202	21	11	110	5
XENOPHON	499	473	503	213	250	133	258	307	0	40	367	11	176	180	320	101	136	132	6	8	131	18	11	92	14
HERODOTUS	442	544	220	476	105	908	52	45	40	0	57	152	39	34	32	217	49	51	104	109	23	105	105	214	499
AUGUSTINE	444	423	582	217	394	141	314	297	367	57	0	5	245	243	398	145	210	122	6	19	230	19	11	133	5
KANT-I	391	595	28	85	24	173	11	6	11	152	5	0	14	20	11	22	13	20	552	372	2	382	752	8	201
AESCHYLUS	385	317	276	118	400	86	475	270	176	39	245	14	0	391	166	127	220	57	19	32	240	40	18	124	2
THUCYDIDES	371	312	264	125	416	90	491	264	180	34	243	20	391	0	166	137	197	55	34	37	219	51	46	118	3
SOPHOCLES	374	374	447	156	224	62	224	263	320	32	398	11	166	166	0	66	98	75	10	14	121	22	7	65	9
OVID	350	253	444	645	446	257	257	140	101	217	145	22	127	137	66	0	204	53	14	49	152	30	19	924	39
HESIOD	346	271	265	162	513	98	277	193	136	49	210	13	220	197	98	204	0	70	15	20	241	23	15	173	9
DIOGENESI	339	326	314	264	121	81	90	107	132	51	122	20	57	55	75	53	70	0	16	14	57	24	26	45	17
HEIDEGGER-M	317	413	20	46	33	118	13	4	6	104	6	552	19	34	10	14	15	16	0	776	9	525	453	9	120
DERRIDA-J	316	290	25	59	65	118	24	11	8	109	19	372	32	37	14	49	20	14	776	0	12	532	346	38	44
NIETZSCHE-F	306	229	265	115	404	59	307	202	131	23	230	2	240	219	121	152	241	57	9	12	0	18	8	128	5
PINDAR	308	284	28	58	60	126	51	21	18	105	19	382	40	51	22	30	23	24	525	532	18	0	346	29	60
HEGEL-GWF	279	383	27	61	35	112	12	11	11	105	11	752	18	46	7	19	15	26	453	346	8	346	0	13	114
VERGIL	279	223	372	615	568	266	207	110	92	214	133	8	124	118	65	924	173	45	9	38	128	29	13	0	38
AQUINAS-T	269	744	38	142	20	405	5	5	14	499	5	201	2	3	9	39	9	17	120	44	5	60	114	38	0

Fig. 4. Matrix of author co-citation data underlying Plato map.

The AuthorWeb PFNET is part of a computer interface that allows users to retrieve the articles that cite the mapped authors and to learn the specific works of theirs that are being cited [6,19]. In another system, maps could be made of co-cited *works*, but AuthorWeb maps of co-cited oeuvres are also useful exploratory tools, especially since one needs to know only an author's name to generate them.

To initiate the AuthorWeb map in Fig. 3, the user enters the name Plato as a seed. The system instantly lists the 24 other names most frequently co-cited with Plato, along with their co-citation counts. These names are "the core of the core" of Plato's bibliogram as a co-cited author. Were the entire bibliogram retrieved, it would have scatter all the way down to authors with whom Plato is co-cited once. The system is intentionally limited to "the seed plus 24" for uncluttered displays.

The user hits a button to create the PFNET, which is the default map [31] and, again, takes only a second or two. Figure 4 has the matrix of co-citation counts from which Plato's PFNET was made; his bibliogram appears in the boxed column at left. (Some near ties are slightly out of order.) Once the seed's top co-citees are retrieved, AuthorWeb pairs *every author in the 24* with *every other author in the 24* and gets a co-citation count for each pair. These are the non-boxed counts in Fig. 4. The relation of the author pairs is undirected (e.g., Aristotle-Cicero has the same count as Cicero-Aristotle). The entries in corresponding columns and rows are thus identical, and the matrix is symmetric above and below the diagonal. AuthorWeb uses counts from only one of its halves. The zero-filled diagonal represents each author singly and is not used in computation. (The matrix here is merely expository; it is not seen by users.)

A bibliogram like Plato's *could* be created for any of his 24 co-citees, and the non-boxed counts are from these potential bibliograms. But note that ranking Plato's co-citees by their counts *un*ranks the counts of everyone else. Moreover, counts for the non-seed authors can come from anywhere in their bibliograms; for example, those of Aquinas with several other authors are from low in their scatter zones.

The PFNET algorithm drastically simplifies the information in the matrix to form the map in Fig. 3. That is, it draws links only between author-pairs with the *highest* (or tied highest) co-citation counts. All 24 of the authors have high counts with Plato; otherwise they would not be mapped. But only five of them, for instance, have their *highest* counts with Plato, and they are the names directly linked to his – Aristotle, Aristophanes, the Bible (a title cited like an author), Homer, and Diogenes Laertius (whose name AHCI truncates). In a contrasting example, Euripides, Pindar, and Hesiod have their highest counts with Homer.

The reader may be able to interpret particular links such as these. But the PFNET algorithm explains why the map makes sense *overall* – the author groups radiating from Homer, from Plutarch, from Kant, and so on. Evidently the *highest* counts in the matrix reflect associations that are *very easy to make* – both for the citers during 1988–97 and, hopefully, for present readers (cf. [31,35,36]). It is much easier to associate, say, Aquinas with Aristotle than with Xenophon; or Heidegger with Kant than with Cicero; or Augustine with the Bible than with Hegel. We have seen similar effects in the other tables in this paper.

A student well-read in the humanities might object that the associations in the map tell us nothing new. There are two answers to this. First, these are the oeuvres that scholars have co-cited most frequently with Plato's *in the literature*. Until his bibliogram is seen, it is not a foregone conclusion what names will be in it and in what order. Second, by extracting information from the matrix, the PFNET arranges the names in meaningful groups that the one-dimensional bibliogram does not. The ancients and the moderns are coherently divided, and the ancients are linked in ways that make them quicker to grasp. The PFNET algorithm does not know, for instance, who Aeschylus and Sophocles are; what it "knows" is that the counts for their letter-strings are highest with the Euripides-string. As users, we see that the Greek tragic playwrights have been felicitously joined. But what the map actually says is that scholars cite Aeschylus and Sophocles with Euripides even more frequently than they do with Plato.

10. Conclusion

The bibliograms in this paper are unified by the claim that their *cores* consist of terms that term-users found easy to associate with the seed; *scatter*, of terms that were less easy. For people examining bibliograms after they are formed, core terms tend to be relatively interpretable; scatter terms, less so. Terms that are opaque to

domain outsiders may be transparent to insiders. Thus we have a psychological explanation of standard bibliometric distributions – one that ties together phenomena that might have seemed unrelated.

This explanation may be briefly contrasted with the probabilistic explanation of core-and-scatter – i.e., power-law – phenomena. The network scientist Mark Newman writes (p. 341): "Thus the probability of a city gaining a new member is proportional to the number already there; the probability of a paper getting a new citation is proportional to the number it already has. In many cases this seems like a natural process. For example, a paper that already has many citations is more likely to be discovered during a literature search and hence more likely to be cited again" [20]. A common way of describing this process is "the rich get richer." Another network scientist, Albert-László Barabási writes (p. 511): "... a newly created Web page will be more likely to include links to well-known popular documents with already-high connectivity, and a new manuscript is more likely to cite a well-known and thus much-cited paper than its less-cited and consequently less-known peer" [3]. He calls the process "preferential attachment." The information scientist Derek Price referred to it earlier as "cumulative advantage" [22].

These count-based, content-neutral explanations of core-and-scatter phenomena are broadly compatible with term-based, content-laden explanations. Newman, Barabási, and Price have backgrounds in physics, and, for them, bibliometric data are best explained through mathematical modeling of impersonal forces. My goal is rather to present verbal and psychological evidence of human agency – even of named individuals – in the creation of bibliometric data [31,35,36]. This does not contradict the mathematical approach; it complements it. An ideal course in bibliometrics would combine both approaches.

Until 2013, a wide variety of full bibliograms could be obtained in Dialog, a search service now extinct. One simply entered a seed term into a Dialog database (e.g., ERIC, INSPEC, Scisearch) and asked the Rank command to list, in descending order, all the terms that co-occurred with it. At present, a more limited set of bibliograms can be generated in the Web of Science, Scopus, Google Scholar (including its Scholarometer and Publish or Perish interfaces), and WorldCat. They should definitely be used in teaching.

As an exercise, for example, the set of papers by a seed author can be retrieved in the Web of Science. The papers can be then sorted by frequency of citation high to low and their bibliographic details examined. If a Citation Report is requested, the high-to-low citation counts are further broken down by year. The report also gives summary statistics and an h-index for the author. The Analyze Results module, already mentioned, can generate ranked lists of other kinds of terms that co-occur with the seed, down to the "onesies." However, the proper steps for various outputs must be learned, and the seed's name must be properly disambiguated.

Unfortunately, bibliograms in full are frequently so long they cannot be reproduced in their entirety. But *extracts* from them, such as are presented here, are a fresh way to introduce bibliometrics to relatively non-technical audiences. In my view, the

best way to engage them is to discuss the verbal side of bibliograms – actual words and names – *along with* the frequency counts and the measures based on them. Let them look at the textual data in the fields of records and in tables. Otherwise, the numeric measures tend to float as disconnected abstractions.

References

[1] F. Åström and J. Hansson, How implementation of bibliometric practice affects the role of academic libraries, *Journal of Librarianship and Information Science* **45** (2012), 316–322.

[2] R. Ball and D. Tunger, Bibliometric analysis – A new business area for information professionals in libraries? Support for scientific research by perception and trend analysis, *Scientometrics* **66** (2006), 561–577.

[3] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, *Science* **286** (1999), 509–512.

[4] M. Bladek, Bibliometrics services and the academic library: Meeting the emerging needs of the campus community, *College & Undergraduate Libraries* **21** (2014), 330–344.

[5] S.C. Bradford, Sources of information on specific subjects, *Engineering* **137** (1934), 85–86.

[6] J.W. Buzydlowski, H.D. White and X. Lin, Term co-occurrence analysis as an interface for digital libraries, in: *Visual Interfaces to Digital Libraries: Motivation, Utilization, and Socio-technical Challenges,* K. Börner and C. Chen, eds, *Lecture Notes in Computer Science* **2539** (2003), 133–144.

[7] P.F. Cole, A new look at reference scattering, *Journal of Documentation* **18** (1962), 58–64.

[8] S. Corrall, M.A. Kennan and W. Afzal, Bibliometrics and research data management services: Emerging trends in library support for research, *Library Trends* **61** (2013), 636–674.

[9] N. De Bellis, *Bibliometrics and Citation Analysis*, Scarecrow Press, Lanham MD, 2009.

[10] M.C. Drott, Bradford's law: Theory, empiricism and the gaps between, *Library Trends* **30** (1981), 41–52.

[11] R. Drummond and R. Wartho, RIMS: The Research Impact Measurement Service at the University of New South Wales, *Australian Academic & Research Libraries* **40** (2009), 76–87.

[12] Edinburgh Word Association Thesaurus, n.d. http://www.eat.rl.ac.uk/.

[13] L. Egghe, The Hirsch Index and related impact measures, *Annual Review of Information Science and Technology* **44** (2010), 65–114.

[14] L. Egghe and R. Rousseau, *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science*, Elsevier, Amsterdam, the Netherlands.

[15] C. Gumpenberger, M. Wieland and J. Gorraiz, Bibliometric practices and activities at the University of Vienna, *Library Management* **33** (2012), 174–183.

[16] B. Hjørland and J. Nicolaisen, Bradford's law of scattering: Ambiguities in the concept of "subject," in: *Context: Nature, Impact, and Role*, F. Crestani and I. Ruthven, eds, *Lecture Notes in Computer Science* **3507** (2005), 96–106.

[17] R. Kear and D. Colbert-Lewis, Citation searching and bibliometric measures; Resources for ranking and tracking, *College & Research Library News* (September 2011), pp. 470–474. (See also Kear's resource guide at http://pitt.libguides.com/bibliometrics.)

[18] M.A. Kennan, S. Corrall and W. Afzal, "Making space" in practice and education: Research support services in academic libraries, *Library Management* **35** (2014), 666–683.

[19] X. Lin, H.D. White and J. Buzydlowski, Real-time author co-citation mapping for online searching, *Information Processing and Management* **39** (2003), 689–706.

[20] M.E.J. Newman, Power laws, Pareto distributions and Zipf's law, *Contemporary Physics* **46** (2005), 323–351.

[21] M.L. Pao, *Concepts of Information Retrieval,* Libraries Unlimited, Littleton, Colorado, 1989.

[22] D.J. deS. Price, A general theory of bibliometric and other cumulative advantage processes, *Journal of the American Society for Information Science* **27** (1976), 292–306.

[23] A. Regolini, E. Gentilini, M.-P. Baligand and E. Jannès-Ober, "Sustainable management" of commercial electronic research resources and of its use in bibliometrics, *Library Management* **34** (2013), 31–39.
[24] R.W. Schvaneveldt, ed., *Pathfinder Associative Networks; Studies in Knowledge Organization,* Ablex, Norwood, NJ, 1990.
[25] A. Strotmann and D. Zhao, An 80/20 data quality law for professional scientometrics? *Proceedings of the International Society for Scientometrics and Informetrics* (2015), 1218–1219. http://www.issi2015.org/en/Proceedings-of-ISSI-2015.html.
[26] J.M. Swales, *Genre Analysis; English in Academic and Research Settings*, Cambridge University Press, 1990.
[27] J.M. Swales, *Research Genres; Explorations and Applications*, Cambridge University Press, 2004.
[28] H.D. White, Author-centered bibliometrics through CAMEOs: Characterizations automatically made and edited online, *Scientometrics* **51** (2001), 607–637.
[29] H.D. White, Authors as persons and authors as bundles of words, in: *Theories of Informetrics and Scholarly Communication*, C.R. Sugimoto, ed., De Gruyter Mouton, Berlin, 2016, in press.
[30] H.D. White, Citation analysis, in: *Encyclopedia of Library and Information Science*, 3d. ed., M.J. Bates and M.N. Maack, eds., Taylor and Francis, New York, 2010, pp. 1012–1026.
[31] H.D. White, Co-cited author retrieval and relevance theory: Examples from the humanities, *Scientometrics* **102** (2015), 2275–2299.
[32] H.D. White, Computing a curriculum: Descriptor-based domain analysis for educators, *Information Processing & Management* **37** (2001), 91–117.
[33] H.D. White, Combining bibliometrics, information retrieval, and relevance theory: Part 2. Implications for information science, *Journal of the American Society for Information Science and Technology* **58** (2007), 583–605.
[34] H.D. White, On extending informetrics: An opinion paper, *Proceedings of the International Society for Scientometrics and Informetrics* **2** (2005), 442–449.
[35] H.D. White, Relevance theory and citations, *Journal of Pragmatics* **43** (2011), 3345–3361.
[36] H.D. White, Some new tests of relevance theory in information science, *Scientometrics* **83** (2010), 653–667.
[37] H.D. White, S.K. Boell, H. Yu, M. Davis, C.S. Wilson and F.T.H. Cole, Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences, *Journal of the American Society for Information Science and Technology* **60** (2009), 1083–1096.
[38] H.D. White and K.W. McCain, Bibliometrics, *Annual Review of Information Science and Technology* **24** (1989), 119–186.
[39] D. Wilson, *Pragmatic Theory*, a course of the Department of Phonetics and Linguistics, University College London, 2007.
[40] D. Zhao, Bibliometrics and LIS education: How do they fit together? *Proceedings of the American Society for Information Science* **48** (2011), 1–4.