Chapter 3

Bibliometric Delineation of Scientific Fields – April 27, 2018

Michel Zitt, Alain Lelu, Martine Cadot, Guillaume Cabanac

Abstract. Delineation of scientific domains (fields, areas of science) is a prior task in bibliometric studies at the meso-level, far from straightforward in domains with high multidisciplinarity, variety and instability. The *Context* section shows the connection of delineation problem to the question of disciplines vs. invisible colleges, through three combinable models: ready-made classifications of science, classical information retrieval searches, mapping and clustering. They differ in the role and modalities of supervision. The *Tools* section sketches various bibliometric techniques on the background of information retrieval, data analysis, network theory, showing both their power and their limitations in delineation processes. The role and modalities of supervision are emphasized. The section *Multiple Networks and Hybridization* addresses the comparison and combination of bibliometric networks (actors, texts, citations) and the various ways of hybridization. In the concluding section, typical protocols and further questions are proposed.

3.1 Introduction

Collecting literature that is both relevant and specific to a domain is a preliminary step of many scientometric studies: description of strategic fields such as nanosciences, genomics and proteomics, environmental sciences; research monitoring and international benchmarks; science communities analyses. Although our focus here is on the intermediate levels, informally described in such terms as areas, specialties, subfields, fields, sub-disciplines... this subject is connected to general science classification and, at the other end of the range, to narrow topic search.

In Sect. 3.2 "Context" we place delineation at the crossroad of two concepts:

the first one is "disciplinarity" (what is a scientific discipline?), which crystallizes various dimensions of scientific activity in epistemology and sociology. The second one is "invisible colleges" in resonance with the core of bibliometrics, the study of networks created explicitly or implicitly by publishing actors. From this point of view, domains of science can be viewed as generalized form of invisible colleges, sometimes in the form of relatively dense and segregated areas — at some scale. In other cases however, the structure is less clear and bounded, with high levels of both internal diversity and external connections and overlaps. Given a target domain, its expected diversity, inter-disciplinarity and instability are challenging issues. We outline the main approaches to delineation: external formalized resources, such as science classifications; ad hoc Information Retrieval (IR) search; network exploration resources (clustering-mapping).

Sect. 3.3 "Tools" is devoted to the main approaches in domain delineation, IR search and science clustering-mapping, when off the shelf classification are not sufficient. Both take root in the information networks of science, but start from different vantage points, with some simplification: ex ante heavy supervision for IR search, typically with bottom-up ad hoc queries; ex-post supervision for bibliometric mapping, with top-down pruning. In difficult cases, these approaches appear complementary, often within multistep protocols. As a result of the complex structure and massive overlaps of aspects of science, of the multiple bibliometric networks involved, of the multiple points of view, the frontiers are far from unique at a given scale of observation. The experts' supervision process is a key element. Its organization depends on the studies context and demand, to reach decision through confrontation and negotiation, especially in high stakes contexts. Beforehand, we shall briefly address the toolbox of data analysis methods for clustering- mapping purposes.

Sect. 3.4 "Multiple Networks and Hybridization" zooms in on the multi-network approach for delineation tasks, stemming from pragmatic practices of information retrieval and bibliometrics. The main networks are actor's graphs and other relations connected with invisible colleges based on documents and their main attributes, texts and citations. Other scientometric networks (teaching, funding, science social networks, etc.) offer potential resources. The hybridization covers a wide scope of forms. There is a strong indication that multi-network methods improve the IR performance and offer a richer substance to experts/users discussions.

3.2 Context

3.2.1 Background: Disciplinarity and Invisible Colleges

Generally speaking there is no ground truth basis for defining scientific domains. Given a target domain, assigned by sponsors in broad and sometimes fuzzy terms, delineation is the first stage of a bibliometric study. It is tantamount to a rule of decision involving sponsors/stakeholders, scientists/experts

and bibliometricians on extraction of the relevant literature. Delineation also matters as research communities are an object of science sociology as well as a playground for network theoreticians.

The delineation of scientific domains should be understood in the context of the structure of science and scientific communities, especially through the game between diversity, source of speciation, and interdisciplinarity drive towards reunification. Disciplinarity and "invisible" colleges are two concepts from the sociology of science which symbolise two kinds of communities, the first one more formal and institutional, the second one constructed on informal linkages made visible by bibliometric analysis of science networks. The tradition of epistemology has contributed to highlight the specificity of science by contrast to other conceptions of knowledge. Auguste Comte proposed the first modern classification of science and at the same time condemned the drift of specialization [1], considered a threat to a global understanding of positive science. In reaction both to epistemology and normative Mertonian tradition [2, 3], Kuhn emphasized the role of central paradigms in disciplines at some point of their evolution [4]. The post-Kuhnian social constructivism proceeded along two lines —at times conflicting [5] — of relativist thinking: the "strong programme" (see Barnes et al. [6]) and the no less radical Actor-Network Theory (ANT). The first one was initiated by Barnes and Bloor [7] and flourished in the science studies movement [8, 9]. The ANT also borrowed from Serres ("translation' concept [10]) and from the post-structuralist French Theory (Foucault, Derrida, Bourdieu, Baudrillard), see [11, 12, 13]. These schools of thought emphasize disciplinarity rather than unity. Lenoir [14] notes that *"A major consequence of [social constructivism] has been to foreground the heterogeneity of science."* Disciplines are *"crucial sites where the skills [originating in labs] are assembled"* and *"political institutions that demarcate areas of academic territory, allocate priviledges and responsibilities of expertise, and structure claims on resources"* (pp. 71–72, 82). Bourdieu stressed the importance of personal relationship and "shared habitus." Disciplines exhibit both a strong intellectual structure and a strong organization. The institutional framework, with, in most countries, an integration of research and higher education systems, ensures evaluation and career management. Some communities coin their own jargon, amongst signs of differentiation, and norms and patterns. Potentially, all dimensions of research activity (paradigms and theories, classes of problems, methodology and tools, shared vocabulary, corroboration protocols, construction of scientific facts and interpretation) appear as discipline-informed, with particular tensions between superdisciplines, natural sciences and social sciences & humanities. Scientists discuss, within their own disciplines, the subfield breakdown and the structuring role of particular dimensions, for example research objects in microbiology, versus integration drive [15, 16].

The endless process of specialization and speciation in science, erecting barriers to the mutual understanding of scientists, is partly counteracted by interdisciplinary linkages which maintain and create solidarity between neighbor or remote areas of research. Piaget [17] coined the term trans-disciplinarity as the new paradigm re-engaging with unity of science. A few rearrangements of

large magnitude, such as the movement of convergence between nanosciences, biomedicine, information and cognitive sciences and technologies (NBIC, concept coined by NSF in 2002), tend to reunite distant areas or at least create active zones of overlap.

In contrast with disciplinarity, the concept of "invisible college" in its modern acceptation, popularized by Price & deBeaver [18] and Crane [19], chiefly refers to informal communication networks, personal relationship and possibly interdisciplinary scope. These direct linkages tend to limit the size of the colleges, although no precise limit can be given. Science studies devote a large literature to those informal groups, which exemplify how actors' networks operate at various levels of science [20, 21].

Although more formal expressions emerge from the self-organization of those micro-societies (workshops, conferences, journals), the invisible colleges do not claim the relative stability and the social organization of disciplines. The various communication phenomena of the colleges are revealed by sociological studies or, more superficially but systematically, by analysis of bibliometric networks such as co-authorship, texts relations and citations. The "bibliometric hypothesis" assumes that the latter process mirrors essential aspects of science: the traceable publication activity, in a broad sense, expresses the collective behavior of scientific communities in most relevant aspects (contents and certification, production and structure of knowledge, diffusion and reward, cooperation and self-organization). It does not follow that bibliometrics can easily operationalize all hypotheses [22]. Affiliations can, in the background, connect to the layers of academic institutions or corporate entities. Mentions to funding bodies are increasingly required in articles reporting grant-supported works. These relations, however, as well as personal interactions, generally require extra-bibliometric information. Variants of the invisible colleges in sociology of science are known as epistemic communities, involving scientists and experts with shared convictions and norms [8, 23] and community of practice [24]. The mix of behavior, stakes and power games, in the interaction of virtual colleges and institutions, remains an appealing question. A revival of the interest for delineation studies has been observed at the crossroads of sociology of science and analyses of networks [25, 26].

Disciplinary views, as well as colleges revealed by bibliometrics, lead to different partitions of literature, depending on the vantage points. In particular, bibliometricians can be confronted with conflictual situations when revealed networks and institutional normative perceptions and claims as to the disciplinary structure and boundaries diverge. The exercise of delineation generally consists in reaching some form of consensus, or at least a few consensual alternatives amongst sponsors, stakeholders, experts and scientists. The toolbox contains information retrieval, data analysis and mapping. Bibliometricians act as organisers of experts' supervision, suppliers of quantitative information and facilitators of negotiations (Fig. 3.1).

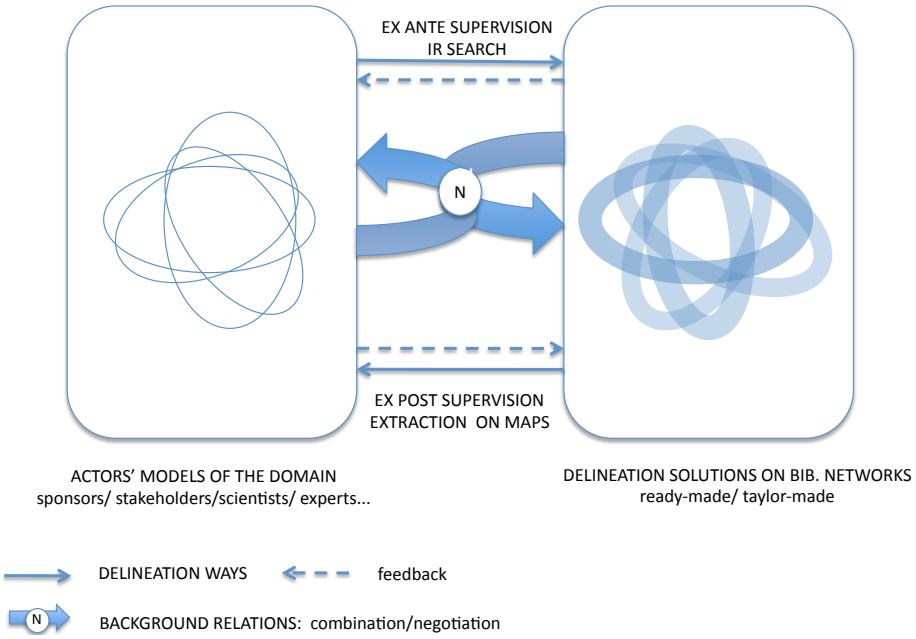


Figure 3.1: Actors' models / bibliometric models. This scheme evokes the interaction between actors' mental or social models of science, disciplines and domains on one hand models from data analyses (clustering-mapping) on bibliometric datasources, based on different methods and networks on the other. The two sides are engaged separately or together in negotiated combinations to reach (almost) consensual views. Two ways of domain delineation are singled out, ad hoc IR search and extraction from maps, with different degrees and moments of supervision. A third way, allowing direct IR search, supposes permanent classification resources.

3.2.2 Operationalization: Three Models of Delineation

In their review of (inter)disciplinarity issues, Sugimoto & Weingart [27] stress that the rich conceptualization of disciplinarity, quite elaborate in sociology and iconic of science diversity, does not imply clear operationalization solutions for defining fields. Scientists' claims and co-optation ("Mathematicians are people who make theorems" with several formulations, including a humoristic one by Alfréd Rényi), university organizations and traditions, epistemology, sociology, bibliometrics offer many entry points. The stakes associated to disciplinary interests and funding, for both scientists and policy makers may interfere with definitions. Introducing the national dimension, for example, shows that the coverage of disciplines is perceived differently in national research systems. Bibliometrics cannot capture the deep socio-cognitive identity of disciplines but contributes to enlighten some of the facets that collective scientists' behavior let appear. The difficulty extends to multidisciplinarity measurement.

In practice, the description of disciplines available in scientific information systems takes the form of classification schemes at some granularity (articles, journals) from a few sources: higher education or research organizations for management and evaluation needs (international bodies or national institutions, for example CNRS in France); schemes associated to databases from academic societies, generally thematic; and/or from publishers or related corporations (Elsevier, ISI/Thomson Reuters/Clarivate Analytics) dedicated to scientific information retrieval.

We term "model A" the principle of these institutional science classifications, which do not chiefly proceed from bibliometrics but from the interaction between scientists and librarians. Subcategories and derived sets offer ready-made delineation solutions. The effect of methodological options, the social construction of disciplines by institutions or scientific societies, with struggles for power and games of interests are unlikely to yield convergence: the various classifications of science available, not necessarily compatible, should be taken with caution. Depending on the update system, they also tend to give a "cold" image of science. Often based on non-overlapping schemes, they tend to handle multidisciplinarity phenomena poorly. Resources associated with classifications in S&T databases which often include various nomenclatures (species, objects) are a distinct advantage. With its limitations this model nevertheless offers a rich substance to bibliometric studies. Since the development of evaluative scientometrics in the 70s, in the wake of Garfield and Narin's works, categories are used as bases for normalization of bibliometric measures, especially citation indicators, but classification-free alternatives exist (see Section 3.3.2). The rigidity of classifications has a advantage, making a virtue out of a necessity, the easy measure of knowledge exchanges between categories over time. Techniques of co-classification [28, 29], co-index or co-word methods (see below) make it possible to transcend the rigidity of the classification scheme.

The concept of virtual college, originally thought as micro or meso-scale communities with informal contours, exchanging in various ways, can be generalized to communities in science networks at any scale. Since the 80s, this

is implicit in most bibliometric studies [30]. Global models of science, either small worlds or self-similar fractal models, are consistent with this perspective. We termed here "model C" this scheme which is the very realm of bibliometrics. Formal and institutional aspects are partly visible through bibliometric networks but need other "scientometric" information on institutional structure of science systems. Bibliometrics and also scientometrics are blind to other networks/relations such as interpersonal networks and to the complete picture of science funding and science society relations. It follows that the delineation of fields in model A, which accounts for complex mixes non totally accessible to bibliometric networks, cannot be retrieved by "model C" approaches. The other way round, model C makes visible implicit structures ignored by the panel of actors involved in model A classification design.

For large academic disciplines, model C merely proposes high level groupings which might emulate the categories "disciplines" from model A and share the same label, however with a quite coarse correspondence. In the practice of model C, large groups receive a sort of "discipline" label through expert supervision. Neither the bibliometric approach nor the model A have the property of uniqueness. Various tests were conducted by external bibliometricians on SCI-WoS (for Science Citation Index of the Web of Science) subject categories, and the agreement is not, usually, that good (e.g., [31]) and the existing ready-made classifications cannot pretend to the status of ground truth or gold standard for domain delineation. Depending on the organization, the clustering-mapping operations often fulfill two needs in bibliometric studies, first helping domain delineation, secondly identifying subdomains/topics within the target. In the absence of ground truth, the challenge of model C is to find trade-offs for reflecting a fractal reality quite difficult to break down, since boundaries are hardly natural except for configurations with clear local minima. They are then subject to optimization with partial information and negotiations [32, 33].

"Model B" based on IR search, borrows from both A and C. In model A, the operationalization of discipline definition and classification relied on heavily supervised schemes, aiming chiefly at information retrieval. Model B shares the same ground, with ad hoc search strategy established by bibliometricians and experts for the needs of the study. Ad hoc search is sometimes necessary in order to go beyond the synthetic views provided by clustering and mapping, and to address analytical questions from users (in terms of theory, methods, objects, interpretation). The default granularity is the document level.

The three models can incorporate a semantic folder. Some indexing and classifications systems provide elaborate structures of indexes and keywords: thesaurus and ontologies (see section 3.2.4). Model B depends on expert's competence and resources of queried databases to coin semantically robust queries. Model C can treat metadata of controlled language, indexes of any kind, as well as natural language texts, and reciprocally shed light, through data/queries treatment, on the revealed semantic structures of universes.

Reflexivity is present under many aspects: scientists are involved in heavy ex ante input in ready-made classifications (model A), in IR ad hoc search (model B) and in softer ad hoc intervention on bibliometric maps (model C).

The supervision/expertise question goes beyond within-community reflexivity, with partners associated to projects: decision-makers and stake-holders and bibliometricians.

Table 3.1 sums up the main features of the three models. They are just archetypes: in practice, blending is the rule. If classical disciplinary classification schemes belong to the first model, the Science Citation Index and variants incorporate bibliometric aspects. Purely bibliometric classifications, if maintained and widely available, give birth to ready-made solutions. In the background of the three models, the progressive approximation of bibliometrics and IR tools, addressed below in the section 3.3 should be kept in mind.

3.2.3 Challenges at the Meso-level

Inter-disciplinarity

Interdisciplinarity is quite an old question, which came to the front of the scene in the early 1970 with the devoted OECD conference and gave rise to an overwhelming literature and programs. The distinction between multi-, inter-, transdisciplinarity formulates various degrees of integration, see [34, 5]. As Choi and Park [35] put it: *"Multidisciplinarity draws on knowledge from different disciplines but stays within their boundaries. Interdisciplinarity analyses, synthesizes and harmonizes links between disciplines into a coordinated and coherent whole."* Jahn et al. [36] detail two interpretations of "transdisciplinarity" in literature. Both make sense in a delineation context. One privileges the science-society relationship: integration between Social Sciences and Humanities (SSH) and natural sciences with the participation of extra-scientific actors, as a response to heavy and controversial socio-scientific problems such as climatic change, genetically modified organisms, medical ethics, etc. The second interpretation considers that transdisciplinarity simply pushes the logic of interdisciplinarity towards integration. Russell et al. [37], cited by Jahn et al. [36], *"emphasize that where interdisciplinarity still relies on disciplinary borders in order to define a common object of research in "areas of overlap (. . .) between disciplines", transdisciplinarity truly "transgresses or transcends [them]".*" Klein [38] and Miller et al. [39] stress the theoretical and problem-solving capability of the transdisciplinary view. Many publications evoke the paradox of multidisciplinarity, a source of radical discoveries, labouring however to convince evaluators in the science reward system. Yegros-Yegros, Rafols & d'Este [40] list a few controversial studies on the topic, and note a specific difficulty for distal transfers. Solomon, Carley & Porter [41] recall that the impact of many multidisciplinary journals is misleading in this respect, since their individual articles are not especially multidisciplinary.

Bibliometric operationalization has to account with those different multi/ inter/ transdisciplinarity forms. 'Multidisciplinarity' involves sustained knowledge exchanges in a roughly stable structure; 'interdisciplinarity', with an organization and systematization nuance, supposes strong exchanges creating some structural strain, between domains overlap and autonomization of merging frac-

Table 3.1: Typical features of the three models for delineating scientific fields

	Model A Ready-made breakdown and tools	Model B Adhoc IR search	Model C Bibliometric Networks
basic concept	science classifications and nomenclatures	union of queries	groupings in science networks, generalizing the concept of invisible colleges
origin	academic societies and database providers. originally little/no input of bibliometrics	publication records (e.g., article metadata) and possibly fulltext	analysis of bibliometric networks from any field in publication records
structure	classification schemes, often hierarchical and hard breakdown (categories: subdisc., fields, specialities, journals, etc.)	categories: if use of structures	networks and clusters/groups at various scales (actors, topics, documents...)
supervision / expertise	heavy ex ante embodied input by scientists, experts, librarians	heavy involvement of scientists, experts, librarians in conception/check of queries	ad hoc softer supervision at various stages (mapping)
data – granularity	richness of added metadata, especially key-words and indexes of objects default granularity: category	all available information, esp. text fields, citation, authors-affiliations default granularity: document	all available information, esp. text fields, citation, authors-affiliations default granularity: cluster
semantic aspects	thesauri, ontologies	structure of queries, use of ready-made resources	latent or explicit dimensions in networks
time features	relative stability of framework, favouring fixed-structure longitudinal analysis, at the expense of tensions in the system between updates	no structural constraint	immediacy and aptitude to dynamic analysis of changing entities

tions; 'transdisciplinarity' paves the way for the autonomy of the overlapping region, within the strong interpretation involvement of SSH and possibly of extra-scientific considerations. Clearly model C is apter than A to depict those forms and their transitions when they occur, rather than waiting for the institutionalization of the emerging structures.

Interdisciplinarity may be outlined at the individual level by co-publications of scholars with different educational or publication backgrounds, by measures of knowledge flows (citations), contents proximity, authors' co-activity or thematic mobility — if such data exist [42]. Other sources include joint programs, joint institutions or labs claiming disciplinary affiliation, generally found in metadata. Most disciplinary databases lagged behind the Garfield SCI model as to the integral mention of all authors' affiliations on an article. The large scope of bibliometric measures of multidisciplinarity was reviewed in many articles, e.g., [27, 43].

In model A the first entry point to multidisciplinary phenomena is the category classification schemes, with measures of knowledge exchanges by citation flows between categories (Pinski & Narin's 1976 seminal work on journal classifications [44], Rinia et al. [45]), transposable to textual proximity (on patents [46]) or authors co-activity. Despite the heavy input of experts in science classification, the delimitation of particular fields varies across information providers and none can be held as a gold standard. It finds its limits in the inertia and often the hard scheme of classes, albeit the derived co-classification and co-index treatments noticed above relax the constraint and instil some of the bibliometric potential of Model C.

Model C is more realistic in depicting the combinatory, flexible, multinetwork relationships in science and the demography of topics. Ignoring disciplinarity as such, it conveys a broader definition of interdisciplinarity, ranging from close to distant connexions, the latter loosely interpretable, in the common acceptation, as interdisciplinary and possibly forerunners of more integrated relations. More generally, the network perspective of model C builds bridges between networks formalization and scientific communities life, leaving open the question of how profoundly the socio-cognitive phenomena are captured. Data analysis methods such as Correspondence Analysis (CA), Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA) addressed below, claim light semantic capabilities at least. Bibliometrics cannot substitute to sociological analysis, which exploits the same tools but goes further with specific surveys. Similarly, it is dependent on computational linguistics and semantic analysis for deep investigations of the knowledge contents. Model C is a potential competitor for offering taxonomies, with recent advances (see Section 3.2.4). It does not follow that dynamics captured by this model are easy to handle: for example, flows variations in a fixed structure (A) read more conveniently than multifaceted structural change (C).

Internal Diversity

Diversity and multidisciplinarity are two facets of a coin. Internal diversity in a delineation process qualifies communities inside the target domain. Fig 3.2B–C expresses the internal diversity of multidisciplinary domains, already striking for nanosciences and massive for proteomics (Fig. 3.2).

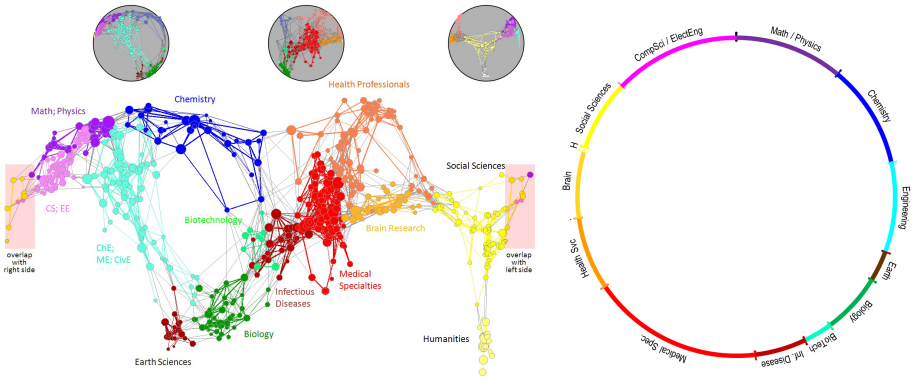
Internal diversity is treated in quite different ways depending on the model. In cluster analysis part of Model C, the balance of internal diversity and external connectivity ("multidisciplinarity" in the looser sense) is part of the mechanism which directly or indirectly rules the formation of groups, with a wide choice of protocols. Many solutions of density measurement are available in clustering or network analysis, with some connection with diversity measures developed in ecology and economics especially. The synthetic Rao index discussed by Stirling [49] combines three measures on forms/categories: variety (number of categories), balance (equality of category populations) and disparity (distance of categories). Delineation through mapping will use smaller scale clusters rather than attempting to capture the target as a whole large-scale cluster. There is no risk of missing large parts of the domain, but the way the different methods conduct the process raises questions about the homogeneity of clusters obtained and the loss of weak signals especially in hard clustering (see Section 3.3).

In model B internal diversity, especially when generated by projected multidisciplinarity, is a threat on recall. Entire subareas may be missed out if the diversity in supervision (panels of experts) does not match the diversity of the domain. Unseen parts will alter the results. In contrast, on pre-recognized areas, model B can be tuned to recover weak signals.

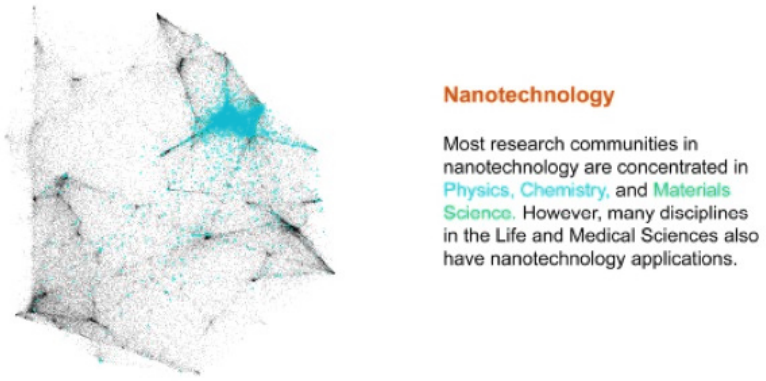
In Model A, the existence of a systemic silence risk particularly depends on how inter-disciplinary bridges are managed.

Unsettlement

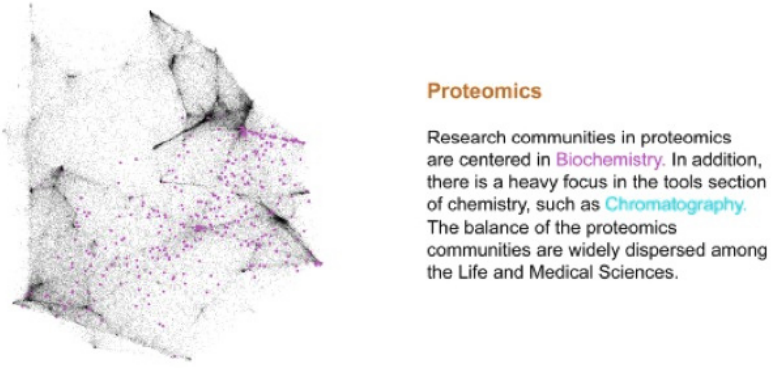
The third challenge of domain delineation lies in the science network dynamics. Conventional model A classifications hardly follow evolutions and need periodic adjustments. The convenience of measures within a fixed structure is paid by structural biases. Bibliometric mapping can translate evolutions in cluster or factor reconfiguration, but the handling of changes in a robust way remains delicate (see Section 3.3). Model B pictures networks, but intuitively, a fast rhythm of reconfiguration in the somewhat chaotic universe of science networks makes it particularly difficult to settle delineation on firm roots. This casts a shadow on the time robustness of the solutions reached on one-shot exercises, but also on the predictive value of extrapolations on longitudinal trends. We go back later to dynamic studies and semantic characterization (see Sect. 3.3.2). Emerging domains seldom embody institutional organization but bear bibliometric signatures of early activity. The difficulty is to capture weak signals with a reasonable immediacy. Fast manifestations of preferential attachment around novel publications, whatever the measure (citations, concept markers or altmetric linkages) are amongst the classical alerts of topic emergence at small scale, to confirm by later local cluster growth.



(A)



(B)



(C)

Figure 3.2: Map of Science and Multidisciplinary projections. Panel A is a world-map type science map from a spherical representation. Panel B and C show hotpots of activity of nanoscience and proteomics projected in a fraction of a global science map. It basically crosses the map holistic picture and an overlay of hits from simple term-queries. Source: Börner et al. [47] – world-type map; Boyack & Klavans [48] – nanoscience and proteomics.

Source Coverage

For memory's sake, the question of data coverage is recurrent in practical bibliometrics and is raised at the delineation stage of any study. The literature on the subject is abundant, conveying different points of view: Hicks [50] first stressed the limitations of both the reference database SCI and the mapping algorithm of co-citation for research policy purposes. Moed's review [51, esp. section 6.2.2] and Van Raan, Van Leeuwen & Visser [52] showed the differential coverage of disciplines by journals in SCI-WoS using references to non-source items. Keeping pace with the growth of visible science is another challenge. The latest UNESCO Science Report estimates that 7.8 million scientists worldwide publish 1.3 million publications a year [53]. SCI-WoS producers proposed new products beginning to fill the gap of books literature, essential to social science & humanities (SSH) and conference proceedings, essential to computer science [54]. The coverage of social science and humanities with issues of publication practices and national biases was addressed in many works, e.g., [55, 56, 57]. This is distinct from the within-discipline approach where an extensive coverage causes instability of indicators due to tails (language biases, national journals biases), to document types or adaptation issues [58, 59, 60, 61]. Former studies' figures are outdated but the basic principles remain.

Large-extension databases with enhanced coverage for IR purposes (modern WoS, Scopus) might require truncation of tails for comparative international studies. The PageRank selection tool limits the noise of a massive extension of sources in Google. However, Google Scholar is not considered a substitute for bibliographic databases for common librarian tasks, but rather a complement especially for coverage extension in long tails [62] with variations amongst disciplines. Same applies to another large bibliographic database: the Microsoft Academic Graph [63, 64, 65]. The lack of transparency in the inclusion process and the lack of tools beyond original ranking (sorting, subject filters) are stressed by Gray et al. [66]. Strong concerns with the quality of bibliographic records were also reported [67, 68]. The coverage of databases has been recently compared by several authors [69, 70], with an extension to alternative sources such as altmetrics: Mendeley, Academia.edu, CiteULike, ResearchGate, Wikipedia, Twitter, etc. [71, 72]. Online personal libraries like Mendeley shed new light on knowledge flows between disciplines through publication records stored together [73] — a kind of co-citation data from readers instead of authors. In addition, these sources, often difficult to qualify properly [74], have been addressed by altmetric studies [42, 75, 76]. The way scientists and the general public communicate about science on (social) media is field-dependent and it is not easy for now to anticipate the complementary role of altmetrics and traditional data in delineation of fields. Altmetric resources can help exploratory and supervision tasks.

In emerging and multidisciplinarity topics that typically justify careful delineation, controversies and conflicting interests are frequent and the importance of transdisciplinary problems makes the issues of sources coverage, experts panel selection and supervision organization more acute.

3.2.4 Ready-made Classifications

Classifications

Table 3.2 below presents some types of science classification valuable in domain delineation. These co-existing classification schemes reflect various perspectives, such as cognitive, administrative, organizational, and qualification-based rationales according to Daraio and Glänzel [77] who stress the difficulties arising when trying to harmonize them.

The first named classifications directly stem from professional expertise of scientists and librarians (pure model A). Some are linked to institutional or national research systems, mainly oriented towards staff management or evaluation, or international instances (UNESCO). More relevant for bibliometric uses are classifications part of complete information systems on S&T literature, proceeding from a few sources: specialized academic societies (CAS, Inspec, Biosis, MathSciNet, Econlit, etc., which usually extend beyond their core discipline) and/or scientific publishers, and patent offices for technology. Classifications are typically hierarchical, complemented by metadata (keywords of various kind, indexes from object nomenclatures: vegetable or chemical species, stellar objects, and so on).

Bibliometrics then entered the competition for science classifications, in contrast with the documentation traditional model involving heavy manpower for indexing individual documents. The prototype is Garfield's SCI/WoS based on the "journal" molecule and a selection tool, the impact factor [81, 82]. The supervision was still heavy in the elaboration of classification, although the Journal Citation Report is a powerful auxiliary for actual bibliometric classification based on journals' citation exchanges [83]. The model of citation index inspired Elsevier's Scopus [84, 85]. The Google Scholar alternative, with a larger scope of less normalized sources, is the extreme case with very little supervision and does not include a classification scheme.

Following Narin's works, several journal classifications were developed (factor analysis in [86], core-periphery clustering in [87]). Many others were proposed over the past decades, some with overlay facilities for positioning activities [88]. Other proposals use prior categories and expert judgments as seeds [89, 90], with reassignment of individual papers. Boyack and Klavans whose experience covers mapping and clustering with several granularity levels (journals, papers) [91] recently reviewed seven journal-level classifications (Elsevier/Scopus ASJC, UCSD, Science-Metrix, ARC, ECOOM, WoS, NSF, JID) and ten article-level classification (five from ISI and CRP, four from MapOfScience, one from CWTS) [92]. The latter authors privilege the concentration of references in review articles ($>$ 100 references) considered as "gold standard" literature, as an accuracy measure (a heavy hypothesis). They conclude in favor of paper-level against journal-level approaches and in favor of direct citations (vs. co-citations or bibliographic coupling) for long term smoothed taxonomies, distinguished from current literature analysis, for which they rank first bibliographic coupling (see below). .

Table 3.2: Science classifications

1. International classifications, often high level. OECD high-level a.k.a. *Frascati Manual*. Fields of Science introduced in 2002. Last revision in 2007. Correspondence table with WoS [78]. 6 major fields were subcategorised.

2. Institutional nomenclature frameworks (ex. CNRS sections[a]). Reflects the vision of the institution and its involvements.

3. Bibliographic databases from science societies. Involve nomenclatures and/or classifications, with a disciplinary focus, sometimes very large (ex. *Chemical Abstracts Service* CAS). Typically based on classical documentation system, with heavy expert input. Another example of classification in Computer Science: *Association of Computer Machinery Classification*[b] 1964-2012.

4. Alternative "ISI model" ISI/Thomson/Thomson-Reuters/Clarivate; Scopus/Scimago Journal Rankings[c] (SJR) as "a publicly available portal that includes the journal and country indicators developed from the information contained in the Scopus database." Firstly used the editorial entity "journal" as the basic molecule, and impact as a principle of selection (see historical account by Garfield [79]). Extensions at a more detailed level. The balance expertise/bibliometrics to design subject categories is unclear (see WoS notices on the topic and [80, p. 1113]). Gives a one- or multi-level hierarchy of groups. The database offers both non-overlapping schemes (Essential indicators) and overlapping schemes (SCI-WoS).

5. Bibliometric mapping classifications, either at the journal or the document level: taylor-made maps potentially usable as permanent resources for public purposes.

[a] http://www.cnrs.fr/comitenational/english/section_acc.htm

[b] http://www.acm.org/about/class

[c] http://www.scimagojr.com/aboutus.php

Those developments mark a new deal in the competition between institutional classification and bibliometric approaches for long-term classifications of science. It is not clear, however, whether the variety of classifications from bibliometric research, not always publicly available, can supersede the quasi-standards of SCI type for current use in bibliometric studies. High-quality delineation of fields cannot solely rely on journal-level granularity, and this is still more conspicuous for emerging and complex domains.

Semantic Resources

Science institutions and database producers have a continuous tradition of maintenance of linguistic and semantic resources, in relation with document indexing. The best known is probably the MESH (National Library of Medicine) used in Medline/PubMed. INSPEC, CAS and now PLOS offer such resources. Controlled vocabulary and indexes, archetypal tools of traditional IR search were also the main support of new co-word analysis in the 80s. A revival of controlled vocabulary and linguistic resources is observed in recent works, associated to the description of scholarly documents [93] and bibliometric mapping [94]. We shall return to the role of statistical tools in the shaping of semantic resources.

3.2.5 Conclusion

Science, seen through scientific networks, is highly connected, including through long-range links reflecting interdisciplinary relations of many kinds. Global maps of science, with the usual reservation on methods settings and artefacts, display a kind of continuity of clouds along preferential directions (Fig. 3.2 (C), from [47]). The extension of domains has to be pragmatically limited by IR trade-off with the help, in absence of ground truth, of more or less heavy supervision. Three models of delineation appear: ready-made delimitation in databases, rather limited and rigid as is, but prone to creative diversions from strict model A (co-classification, etc.); model B, ad hoc search strategies combining several types of information; model C, by extraction of the field from a more extended map, regional or global.

Networks of science may locally show cases of domains ideal for trivial delineation: a perfect correspondence between the target and ready-made categories, or insulated continents surrounded by sea. Such domains will not require sophisticated delineation. This is the exception not the rule.

Areas such such as environmental studies nanosciences, biomedicine, information and cognitive sciences and technologies (converging NBIC, concept coined by NSF in 2002) exhibit both internal diversity and strong multidisciplinary connections. Commissioned studies often target emerging and/or high-tech strategic domains which witness "science in action" prone to socio-scientific controversies à la Latour. These areas combine high levels of instability and interdisciplinarity. As to transdisciplinarity, the question arises of whether to include SSH and alternative sources in datasources and panels experts.

3.3 Tools: IR and Bibliometrics

This section focuses on some technical approaches of the delineation problem: information retrieval and bibliometric mapping. They share the same basic objects and networks, chiefly actors and affiliations, publication supports, textual elements and citation relations. Although general principles of bibliometric relations studies are quite established, new techniques from data analysis and network analysis, including fast graph clustering, open new avenues for achieving delineation tasks on big data at the fine-grain level. The quality of results remains an open issue. Domain delineation confronts or combines the three approaches previously stated: ready-made categories (model A) are seldom sufficient; we shall envision ad hoc IR Search (model B) with an occasional complement of ready-made categories; and on bibliometric processes of mapping/clustering along model C.

3.3.1 IR Term Search

The question of delineation spontaneously calls for a response in terms of information retrieval search. The only particularity is the scale of the search or more exactly, as mentioned before, the diversity expected in large domains, which is particularly demanding for the "a priori" framework of information search. The verbal description of the domain requires, beforehand, an intellectual model of the area. In addition to the methodological background brought by IR models, a broad range of search techniques address delineation issues:

- Ready-made solutions in the most favourable cases, with previously embodied expertise, sketched above.
- Search strategies of various levels of complexity, also depending on the type of data, relying on expert's sayings.
- Multistep protocols: query expansion, combination with bibliometric mapping.

IR models are outside the scope of this chapter. In the tools section below, we recall some of the techniques shared by IR and bibliometrics, especially the vector-space derived models.

IR tradeoff at the meso-level

The recall-precision trade-off is particularly difficult to reach at the meso-level of domains exhibiting high diversity. Generic terms (say the "nano" prefix if we wish to target nanosciences and technology) present an obvious risk on precision. A collection of narrower queries (such as "self-assembly," "quantum dots," etc.) is expected to achieve much better precision. In the simpler Boolean model, this will privilege the Union operator of subareas descriptors (examples on nanoscience [33, 95, 96]). However, nothing guarantees a goodness of coverage of the whole area by this bottom-up process. An a priori supervision of the

process by a panel of experts is required, but the experts' specialization bias, especially in diverse and controversial areas, generates a risk of silence. Similar risks are met in the selection of training sets in learning processes. Another shortcoming is the time-consuming nature of supervision, again worsened by the diversity and multi-disciplinarity of the domain. A light mapping stage beforehand may reduce the risk of missing subareas. As mentioned above, focused IR searches are, in contrast, able to retrieve weak signals lost in hard clustering.

Poly-representation and Pragmatism

Scientific texts contain rich information, most of it made searchable in the digital era. Pragmatically, all searchable parts of a bibliographic record, data or metadata are candidates for delineating domains: words n-grams in titles, abstracts and full texts; authors, affiliations, date, journal or book, citations, acknowledgements, transformed data (classification codes, index, controlled vocabulary, related papers...) depending on the database. These various elements exhibit quite different properties. In theoretical terms, the variety of networks associated to these elements are one aspect of the "poly-representation" of scientific literature [97]. We go back to this question later (Sect. 3.3.2). A specific advantage of lexical search is the easy understanding of queries — whereas other elements (aggregated elements such as journals; citations) are more indirect. However, the ambiguity of natural language reduces this advantage.

Bibliometric literature is packed with examples of pragmatic delineation of domains based on IR search. By and large, apart from ready-made schemes when available (indexes, classification codes), a typical exploration combines a search for specialized journals if any, and a lexical search in complement. At times, an author-affiliation entry is used, especially in connection with citation data. Bradford and Lotka ranked lists are therefore good auxiliaries, with evident precautions on journals or authors' degree of specialization.

Granularity

We noted above that some ready-made classifications such as SCI scheme (journals or journal issues) are essentially based on full journals — or journal sections. These ready-made categories very seldom fit the needs of targeted studies. Instead, ad-hoc groupings of selected journals relatively easy to set up with the help of experts, are a convenient starting base within a Bradfordian logic. The journal level presents obvious advantages. Journals exhibit a relative stability in the medium term; they are institutionalized centres of power through gatekeeping, and a (controversial) evaluation entity in the impact factor tradition.

However, the journal level is problematic for delineation studies. Journals whose specialization is such that they indisputably belong to the target domain, can be taken as a whole, but of course target domain literature are rarely covered by specialized journals only, and investigations should be extended to moderately or heavily multidisciplinary sources. Conditions of diversity and multi-disciplinarity — which prevail in the targets of studies where elaborate

delineation is worthwhile — hinders efficiency of global Bradford/Lotka based selections, with problems of normalization (see also [98]). We go back to these issues in the sub-section devoted to clustering and mapping.

To conclude on this part, the IR resources in scientific texts, data and metadata, suggest a poly-representation of scientific information (cognitive model [97]), which is akin to the multi-network representation of the scientific universe. Ingwersen & Järvelin [99, p. 19] propose a typology of IR models and the perspective of the "cognitive actor." IR protocols generally involve multistep approaches, with various core-periphery schemes (see below). In conventional search, heavy ex-ante supervision is needed for covering the variety of domains, ideally with good analytic/semantic capability. In the absence of gold standard, proxy measures of relevance are needed.

Multistep Process

Multistep processes, possibly associated with combination of various bibliometric attributes, are run-of-the-mill procedures (see for example [32]).

Core-periphery rationale is common, in accordance with the selective power of concentration laws, both in IR and bibliometrics (journal cores in [100], co-citation cores in [101], H-core in [102], emerging topics in [103]). For example, working on highly cited objects — authors, journals or articles — gives a set of reasonable size, amenable to further expansion with enhanced recall. Cores inspired from Price law on Lotka distributions or from H-index application are helpful. Proxies such as seeds obtained from initial high-precision search stages can do as well. The core or seed expansion process is global or cluster-based. The risk of core-periphery schemes, by and large favourable to robustness, is to miss lateral or emerging signals. This may need some input of dynamic characterization of hotspots at fine granularity level.

A parent method is bibliometric expansion on citations, which also uses information from a first run (set of documents retrieved by a search formula or a prior top cited selection, considered as the core) to enhance the recall through the citation connections, typically operating at the document level with or without clustering/mapping step. In this line the Lex+Cite approach mentioned in Section 3 relies on a default global expansion, rather than a cluster-based one, to limit the risk of an exclusive focus on cluster level signals that would miss across-network bridges.

Query expansion by adaptive search is in the same line. Interactive retrieval with relevance feedback identifies the terms, isolated or associated (co-occurrences), specifically present in the most relevant documents retrieved according to various measures [104, 105, 106]. An efficient but heavy process consists in submitting the output of a search stage to a data analysis/topic modeling, able to reconstruct the probable structure likely to have generated the data. By providing information on the linguistic context — also citation, authoring context, etc. — they in turn help to improve the search formulas by a kind of retro-querying. This ranges from simple synonyms detection to construction of topics, orthogonal or not, suggesting the rephrasing of queries.

Variants of itemset mining uncovering association rules [107, with earlier forerunners] are promising in this respect (see below). Evaluation of output from unsupervised stages can also call for a manual improvement of queries.

Delineation protocols may also use the seed as a training set for learning algorithms. A difference is that core-periphery schemes usually rely on the selective power of bibliometric laws, whereas the training set might be extracted on various sampling methods, provided that the seed does not miss the variety of the target. As Big Data grows bigger, "semi-supervised" approaches are gaining popularity in the machine learning community. This recent approach should prove attractive in the bibliometrics community, as considerable interest seems raising for linking metadata groups and algorithmically-defined communities [108].

To conclude on this part, whilst typical IR search relies on an "a priori" understanding of the field, multistep schemes involve stages of data analyses quite close to bibliometric mapping practices, the topic of the next subsection. IR and bibliometrics share roots and features, which soften the differences: adaptive loops, learning processes, seed-expansion and core-periphery schemes. Bibliographic coupling, at the very origin of bibliometric mapping, came from the IR community [109] and the "clustering hypothesis" about relevant vs. non relevant documents [110] voice the common interests of IR and bibliometrics, beyond the background methodology of information models (Boolean, vector-space or probabilistic) and general frameworks such as the above-mentioned cognitive model. The tightening of bibliometrics-IR relations has been echoed in a series of workshops and in dedicated issues of *Scientometrics* ([111, 112], see also [113] for a focus on domain delineation) and in the *International Journal on Digital Libraries* [114].

3.3.2 Clustering and Mapping

In contrast with conventional IR search, bibliometric mapping starts at a larger extension level than the targeted domain. This broad landscape, typically built by unsupervised methods, is scrutinized by experts to rule out irrelevant areas. The supervision task is limited to the "post-mapping" stage. This is in principle less demanding than the a priori conception of a search formulation or of a training set. The default solution is a zoomable general or regional map of science, with availability and cost constraints. The alternative is the construction of a limited overset including almost certainly the anticipated domain, using general search set for massive recall, an operation much lighter than the set-up of a precise search formula. In terms of scale, the final result is tantamount to the outcome of a top-down elimination process, although the selection modalities are diverse. There is currently a great interest in delineation through mapping. IR and mapping are complementary in various ways. Firstly, we shortly describe the data analysis toolbox, before addressing the main bibliometric applications and a few problematic points.

Background Toolbox

The data structure of matrices in the standard bibliometric model allows scholars to mobilize the large scope of automatic clustering, factor/postfactor methods and graph analysis. Classical methods of clustering and factor analysis keep going in bibliometrics, but in the last decade(s) novel methods came of age, more computer-efficient and fit for big data, an advantage for mapping science and delineating large domains. Starting with bibliometric data of the standard model and some metrics of proximity or distances, clustering and community detection methods produce groups. Element are mapped using various dimension reduction algorithms. Factor methods produce groups through clustering applied to factor loadings, with an integrated 2D or 3D display when just two or three factors are needed in the analysis.

A major driving force of bibliometric methodology is the general network theory, which took large networks of science, especially collaboration and citation, as iconic objects [115, 116, 117]. Quite a few mechanisms have proposed to explain or generate scale-free networks since Price's cumulative advantage model for citations [118] in the line of Yule and Simon, and later studied in new terms (preferential attachment) by Albert & Barabási [119], see also [120]. These models have some common features with the Watts-Strogatz small worlds model, but also differences empirically testable [121]. Amongst other mechanisms: homophily [122], geographic proximity [123], thematic proximity inferred from linguistic or citation proximity. Börner et al. reviewed a few issues in science dynamics modeling [124]. Of great interest in bibliometrics and especially delineation, community detection algorithms exhibit a general validity beyond "real" social networks, and belong to the general toolbox of mathematical clustering and graph theory — applicable to various markers of scientific activity, document citations, words, altmetric networks, etc. see also [120].

Hundreds of clustering and mapping methods have been designed during a one-century time lapse of uninterrupted research. This section can only provide a basic overview of the main method families, in the perspective of domain delineation. More comprehensive descriptions and references, as well as a basic benchmark of various methods, applied to a sample of textual data, can be found in [125].

Clustering Methods. Although *hierarchical clustering* algorithms sometimes seem old-fashioned because of their computing complexity, $O(n^2)$ in the very best cases, some of them show good performances for relative small universes. For large ones, they can be coupled to beforehand data reduction stages, classical (SAS Fastclus $O(n)$), pre-clustering algorithms for big data (Canopy clustering [126]), or sampling methods. All-science bibliometric maps rather use faster algorithms today, not without limitations however. Discipline-level maps, or simply internal clustering of the domain set at various stages of delineation may still rely on the classical techniques.

Hierarchical ascending algorithms are local, deterministic and produce hard clusters, with a few exceptions (pyramidal classification), properties favorable

to dynamic representations. They do not constrain the number of clusters and provide multiscale view through embedded partitions, with some indication of robustness of forms in scale changes. Most hierarchical descending (divisive) methods are heavier. Hierarchical methods typically rely on ultrametrics, which down-sides, see [125].

Amongst popular methods in bibliometrics are ascending methods: single linkage, average linkage and Ward. Single linkage is relatively fast and exhibits good mathematical properties in relation to spanning trees but produces disastrous chain effects which must be limited in various ways. Ward and especially group average linkage give better results. Group average linkage advocated for bibliometric sets by Zitt & Bassecoulard [127] and used by Boyack & Klavans in various works [128] is slightly biased towards equal variance and is not too sensitive to outliers. Ward is biased towards equal size with a strong sensitivity to outliers. Properties and biases were studied especially by Milligan [129, 130] using Monte Carlo techniques.

Density methods are appealing: deterministic too, local, and as such prone to dynamic representations of publication or citation flows. DBSCAN [131] (for Density-Based Spatial Clustering of Applications with Noise) is the most popular to the point of becoming synonymous with "density clustering." The SAS clustering toolbox includes hierarchical methods with prior density estimation, with good properties towards sampling and ability to capture of elongated or irregular classes. However, this property is disputable in bibliometric uses (see Sect. 3.3.2, cluster shape/properties of clusters). More recently Density Peaks [132] implements an original and graphical semi-automatic procedure for determining the cluster seeds.

Not directly hierarchical is the venerable *K-means clustering* family, still popular, thanks both to its excellent time/memory performance and sensitivity to different cluster densities. A shortcoming of not being deterministic, they converge to local optima of their objective function, depending on their random (or supervised) initialization. In comparative analyses, they are not considered too sensitive to outliers. They optionally allow for soft/fuzzy clusters, and approximate dynamic data-flow analysis.

Factor methods are basically dimension reduction techniques, indirectly linked to the partition problem. A quick-and-dirty heuristics for extracting a limited number k of dominant clusters from k factors consists of assigning each entity to the factor axis which maximizes the mode of its projection, subject to the constraint of a common factor sign for the majority of entities assigned to this cluster — which eliminates few of them in practice. For a more rigorous procedure, see the descending hierarchical clustering method Alceste [133] in the dataspace of Correspondence Analysis. Factor methods rely on the mathematical foundation of Singular Value Decomposition (SVD) of data matrices for reducing dimensionality and filtering noise. The interesting metrics used by Correspondence Analysis (CA [134]) explains the attention from many scholars for half a century for mappings or clusterings limited to a few dominant factor dimensions. Dropping this limit, i.e. taking into account factor spaces with hundreds of dimensions [135], latent semantic analysis (LSA [136]) unblocked

and fostered the integration of semantics in textual applications, in a lighter but more convenient form than handmade ontologies, costly to edit and update.

Hybrid factor/clustering methods, sometimes coined *topic models*, result in representing each cluster as a local, oblique factor, with a progressive scale from core elements to peripheral ones, opened to fuzzy or overlapping interpretations or extensions. Generally powered by the Expectation Minimization algorithm (EM), they converge to local optima, too. Non-negative Matrix Factorization (NMF) and Self-Organizing Maps (SOM) are well-known examples. Axial k-Means (AKM in [137]) has been used in a comparative citations/words bibliometric context (see Sect. 3.4).

Also known as topic models, the *probabilistic models* try to lay solid statistical foundations for their hybrid-looking representation: they explicit generative probabilistic models for the utterance of topics and terms [138]. Probabilistic LSA (pLSA in [139]) and Latent Dirichlet Allocation (LDA in [140]) are the best-known examples, claiming good semantic capabilities. The older Fuzzy C-means Method (FCM) is akin to this family, which uses the EM scheme for converging to local optima of their objective function.

The *graph clustering family*, also known as *network analysis*, or *community detection methods*, does not operate on the raw (entities*descriptors) matrix, as the previous families do, but on the square (entities*entities) similarity matrix, whose visual counterpart is a graph. Most of these methods operate directly on the graph, detecting cliques or relaxed cliques (modal classification), e.g., Louvain [141], InfoMap [142] and Smart Local Moving Algorithm (SLMA in [143]). Some of them operate on the reduced Laplacian space drawn from the graph (spectral clustering [144]). Quite a few comparative studies are available [145, 146, 147].

Note on Deep Neural Networks. While neural networks were somewhat in standby during the 1995–2005 decade, challenged by more manageable mathematical methods, several factors like the pressure of big data availability and progress in hardware (GPU, i.e., Graphics Processing Units) triggered a renewal under the banners "deep neural nets" and "deep learning." Allowing learning by back-propagation of errors in many layers network, they gave form to the dream of knowledge acquisition by growing levels of abstraction: for images, extraction of local features; contours, homogeneous areas, shapes; for written language: characters n-grams, words, words n-grams, expressions/phrases, sentences. Typically, they avoid heavy Natural Language Processing (NLP) pre-processing (parsing, unification, weighting, selection...). These techniques are already widely used in supervised learning, with spectacular progress in automatic translation, face recognition, listening/oral comprehension, with important investment from the largest internet-related companies (e.g., Google, Apple, Facebook, Amazon), especially. As far as informetrics and IR are concerned, the main domain impacted so far is logically the large scale retrieval (e.g., see [148] which uses a robust letter-trigram based word-n-gram representation). There are also some attempts of non-supervised processes for information

retrieval [149].

A promising technique is the Neural Word Embeddings (NWE). Millions of texts now available online make it possible to develop vector representations of words in a semantic space in a more elaborate way than LSA — a method coined "Neural Word Embeddings." For example, the Word2Vec algorithm [150] processes raw texts so as to list billions of words-in-context occurrences (e.g., word + previous word + next word), then factorize [151] the word × context matrix (tens of thousands words, a few hundreds of thousands, or millions unique contexts) and extract some hundreds or thousands semantic and syntactic dimensions. We go back later to NWE semantic capabilities.

Note on the definition of distances. Whether starting from a binary presence/absence matrix or from occurrence or co-occurrence counts, some methods embed a specific weighting scheme, i.e. a metrics, for computing distances, or similarities between items. This is the case of probabilistic models, Correspondence Analysis, and Axial K-Means. Other methods allow for a limited and controlled choice, as aggregative hierarchical methods do. In the case of graph clustering methods, the user may freely choose his preferential distance definition prior to building the adjacency matrix, which adds an extra degree of freedom beyond the choice of the degree of non-linearity, via a threshold value. For word-based matrices, heavier than citation-based ones the methods of k-means family are also making it possible to choose a weighting scheme (Salton's Term Frequency – Inverse Document Frequency (TF·IDF), Okapi-BestMatch25 [152]).

Whereas factor/SVD methods combine metrics and mapping capability, e.g., two-factor planes or 3D displays, at the native granularity level (e.g., document × words), other mapping algorithms may operate on rectangular or on square (distance) matrices of elements or on groups from a clustering stage, or institutional aggregates (journals). Families of mapping techniques rely on various principles: equilibrium between antagonistic forces — repulsion between nodes, attraction alongside edges (e.g., Fruchterman & Reingold algorithm [153], implemented in Gephi [154], alone or combined with clustering (Sandia VxOrd/DrL/OpenOrd [155], CWTS VOS viewer [143]); optimization of diverse functions: projection stress minimization in the case of MDS, with Euclidean distances in the case of metric MDS, a variant of PCA, and other distances or non-linear functions of these distances in the case of non-metric MDS, one of the non-linear unfolding techniques; maximizing inertia in the case of Correspondence analysis, minimizing edge-cuts in a 2D projection plane; or maximizing local edge densities (Pajek [156]).

Itemset Techniques. Itemset techniques are used for describing a data universe in terms of simple procedures, typically Boolean queries with AND, OR and NOT operators. This may be used for building a stable procedural equivalent of data, e.g., for updating a delineation task (like probabilistic factor analyses). It may also be used for query expansion, as mentioned above in

Sect. 3.3.1. The problem amounts to duplicate a reference partition in a new universe: machine learning techniques are basically fit to this problem, and, in the particular context of textual descriptions, itemset techniques. They are akin to generate Boolean queries with AND, OR and NOT operators, for extracting approximations of the delineated domain, within precision and recall limits established in the machine learning phase [107, 157].

A Benchmark. To illustrate the capabilities of these various methods on an example, in the absence of a bibliometric dataset libelled with indisputable "ground truth" classes we turned towards a reference dataset popular in the machine learning community, the Reuters 21 578 ModApté split (The corpus description is available online at http://www.daviddlewis.com/resources/testcollections/rcv1/. The website http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html has made a pre-processed version of this corpus available to the public, as a supplementary material to [158]). Main features are:

- Source: a set of short texts: newswires from press agencies.
- Contents: in its 6-class selection used, the number of texts (∼7,000) and terms (∼4,000) is sufficient with regards to text statistics.
- Class structure considered as ground truth: built by experts, visually glaring in Fig. 3.3: two big classes, one very dense, the other not, and four small classes, two of which are linked together. In this way, two major problems of real-life datasets are addressed: the imbalance between cluster sizes, and between cluster densities.

We challenge 17 clustering/mapping methods to retrieve this class structure. The similarity of their cluster solution to ground truth partition is measured by two indicators, Adjusted Rand Index (ARI [159]) and Normalized Mutual Information (NMI [160]). Results are detailed in [125]. Let us summarize them in a user-oriented view, sorted by number of required parameters: the lesser the better, ideally, facing a bibliometric dataset without prior knowledge, no parameter.

- Two methods of network analysis require no internal parameterization, Louvain and InfoMap. However the similarity matrix generally requires a threshold setting, here fixed to 0.1 in the cosine inter-text similarity matrix. Infomap obtains the best result in term of NMI (0.436 value vs. 0.423), the index considered the best match for human comparison criteria. This value is rather poor, this method does not distinguish Classes 1, 2, 3, 4, and splits Class6.
- Nine methods require one parameter: the three hierarchical clusterings need a level cut parameter, possibly adjusted for 6 resulting clusters, while for CA, NMF, AKM, pLSA, LDA and Spectral Clustering, the number of desired clusters (6) has to be specified. As the latter group converges to

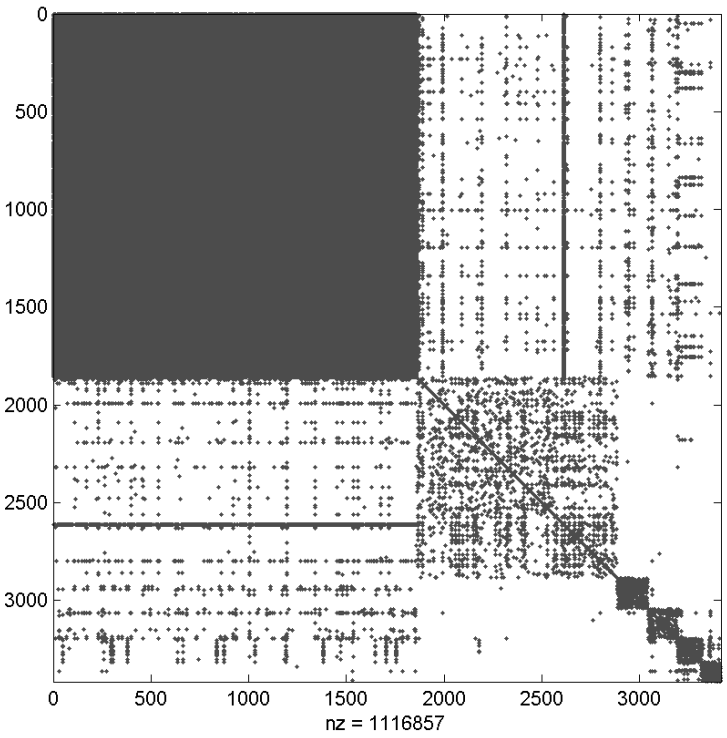


Figure 3.3: Benchmark structure (ground truth). "Spy" plot of the cosines between document vectors of the top6-classes Reuters ModApté split collection. The rows and column ordering is that of the six Reuters classes. Black pixels mean: cosine > 0.5.

local optima, we kept the best results in term of their own objective function out of 20 runs. The indisputable winner is Average link clustering, in both ARI (0.62) and NMI (0.71) terms. The lists of the four following challengers are contrasted: with regard to ARI, first Mac Quitty hierarchical clustering (0.50), then LDA, AKM, CA; with regard to NMI, first AKM (0.51), then Mac Quitty, CA, LDA. If one optimizes ARI all over 20 runs with prior knowledge of the 6-clusters structure — a heroic hypothesis —, Average link clustering still performs best (with a ten-clusters cut, ARI = 0.71, NMI = 0.64) while the followers reach, at best, ARI = 0.55 and NMI = 0.55.

- The last group of methods (ICA, DBSCAN, FCM, Affinity Propagation, SLMA, Density peaks) require at least two parameters, a handicap in absence of prior knowledge of the corpus structure. SLMA obtains the best rating (ARI = 0.60, NMI = 0.55).

Our general conclusion is that one must be very cautious regarding domain delineation resulting from one run of one method. Multiple samplings, if necessary, and level cuts of Average links as well as multiple runs of LDA, AKM and SLMA may help determine core clusters, and possibly continuous gateways between them. Limitations of this benchmark exercise should be kept in mind. It would benefit from tests on different reference datasets: any method can be trapped in particular data structures, and the results cannot be extrapolated without caution. As advocated below, processing multiple sources (lexical, citations, authors...) and investigating the analogies and differences in their results will always prove rewarding. A number of in-depth benchmarking studies are found for hierarchical clustering (Milligan [129, 130] not covering the last techniques), discussing the generation of test data as well as comparisons of algorithms. For community detection, usually taken as a synonym of graph-based clustering rather than clustering of true social (actors) communities, [145] ranked first Infomap, then Louvain and Pott's model approach [161]. Leskovec et al. [146] studied the behavior of algorithms with increasing graph size. Yang & Leskovek [147] add a reflection on the principles of clustering faced to institutional classifications.

Bibliometric mapping

Classical Way. Most classical bibliometric mapping, as well as information retrieval, relies on substantive ("feature") representations of words, words combination, citation, indexes and so forth. Substantive representation implies legibility and interpretation by experts or users, and a condition for bibliometricians or sociologists to check and possibly deconstruct the document linkages. It contrasts with featureless machine representation applicable for example to distances of texts (see below). In contrast, the substantive approach is deepened in semantic studies: ontologies and semantic networks suppose more elaborate investigation of terms relationship. Bibliometric mapping and IR techniques are both a client of ready-made semantic resources, and providers of studies,

supported by data analyses, likely to help the construction of thesauri and ontologies.

The standard bibliometric model starts from the data structure of articles, essentially a series of basic article × attributes matrices, one of these reflexive: article × cited references, where references can also stand as attributes. The derived article × article matrices (e.g., bibliographic coupling, lexical coupling) and elements × elements matrices (e.g., co-word or profiles, co-citation or profiles) cover a wide range of needs. Clusters of words are candidates for conceptual representation, concepts which in turn can index the documents. Likewise, clustering of cited articles reveal intellectual structures and in turn index the citing universe. Basically, the attributes (words from title, abstract, full text; keywords list, indexes — other fields like authoring) are processed in bags of monoterms or multiterms, recognized expressions or word n-grams. Standard bibliometric treatments go rarely further, semantic studies do, for example by using chain modeling of the texts. All these forms allow for control and interpretation of linguistic information.

Assuming that the final purpose is to classify or delineate literature, the access is dual: direct classification of articles after their profile on the structuring elements (words, cited references), or a detour by the structuring items: word profile (especially co-word), citation profile (co-citation), index (or class profile) including co-classification, when applicable. The basics of citation-based mapping were established in the 60s and the 70s: bibliographic coupling [109], chained citations [162], co-citation [101, 163], author co-citation [164], co-classification, etc. The lexical counterpart, with its first technical foundations in Salton's pioneer works [165], was reinvested by English and French social constructivism in the 80s [166, 167, 168] with a stress on local network measures quite in line with the development of social network analysis in that period [169]. In bibliometrics, the true metric approach of text-based classification, Benzécri's correspondence analysis [134], remained confidential. For convenience reasons, many large-scale classifications relied on proximity indexes and MDS or hierarchical single-linkage (ISI co-citation). We return later to word-citation comparison and combination (Sect. 3.4).

Developments. The principles above, mutatis mutandis, are kept in further developments of citation mapping: the approach through citation exchanges, mentioned in Sect. 3.2, assumes predefined entities, journals for example. At the article level, symmetrical linkages between articles, or between structuring elements, are classical: large-scale co-citation (CiteSpace [170]). Glänzel & Czerwon [171] advocated bibliographic coupling. As already mentioned, direct citation linkages clustering, first benchmark for co-citation and coupling in Small's princeps paper [101], is considered as particularly able to reflect long period phenomena [172, 173, 92] but not short-term evolutions. It comes out that the time range picked and the granularity of groupings desired might suggest the choice between the three families of citation methods to reflect structure and changes in science.

From the theoretical point of view, co-citation (respectively co-word) is semantically superior to coupling, by visualizing the structure of the intellectual (cognitive) base, but requires a secondary assignment of current citing literature. Coupling as such, because it by default spares the dual analysis (the cited structure; the lexical content), is semantically poor but bibliographic coupling handles immediacy better than co-citation does. However, this depends on the computer constraints and the settings: the thresholding unavoidable in co-citation analysis drastically reduces weak signals that are counted for in coupling. The dependence of the maximum recall on the threshold of citation and the assignment strength (number of references), in a close field, is modeled in [59]. that are accounted for in coupling. Quite a few authors compared the methods empirically [128] on short time range, [173, 174]. These studies are not always themselves comparable in their criteria, nor are they convergent in their outcome, so that it is difficult to come to a conclusion on this basis alone.

The new data analysis toolbox (fast graph unfolding, topic modeling) gradually pervades large-scale studies. From the domain delineation perspective, a general answer in terms of single best cannot be expected. The benchmark above recalls us that classical methods, apparently outdated in the big data era, still prove quite performing. Let us recall a few issues in clustering/mapping for bibliometric purposes, especially delineation.

A Few Clustering/Mapping Issues

As other decision-support tools, maps in bibliometrics receive contrasted interpretations. In a social constructivist view, maps are mainly viewed as tools of stimulation of socio-cognitive analysis and also as supports of negotiation with/amongst actors. If technicalities are not privileged, there is clear preference for local network maps, preferably lexical or actors-based, connected to socio-cognitive thinking. Bibliometricians and librarians are keener on quantitative properties and retrieval performances. Expectations as to ergonomics, granularity, robustness, clusters properties and semantic depth, largely vary depending on the type of study.

Ergonomics. Maps usage benefits from new displays with interaction facilities. A tremendous variety of mapping methods is available (see [175] although in practice a few efficient solutions prevail. The progress in interfaces (scale zooms, bridges between attributes, interaction with users...) changed the landscape of mapping. If adding clusters features to cluster maps is trivial [176], the systematization of overlay maps by Leydesdorff & Rafols [177] is quite appealing. Since delineation tasks often deals with multidisciplinarity, multiassignments and cluster expansion, various types of cross-representations (see Sect. 3.4) including overlay maps are quite convenient tools for discussion.

Granularity. The granularity considered here is the smallest unit handled. Progress of data analysis allows large-scale work with a fine granularity. Document-level maps are now regularly proposed by Boyack & Klavans (e.g., [91]).

The classical alternative in bibliometrics uses the "journal" molecule instead of publications, with the advantages and shortcomings already discussed. Delineation tasks used to be conducted at the journal level and this convenient solution can be somewhat improved using core-periphery scheme with multidisciplinary qualification [178].The interest of the journal granularity for delineation remains dependent on their specialization profile at the scale considered, quite field-dependent. The best fit to journal approach is found in fields with strong editorial focus, such as Astrophysics, but [179] recalls that the general rule is the superiority of document granularity. At the global science level, journals or even journal categories are an option for sketching great regions [177], with low precision ambitions. In favour of journals, their persistence as institutional entities with slow demography, facilitates longitudinal approaches, again at the expense of precision (see below "Dynamic Clustering"). Granularity does not reduce to the question of journals vs. document level. It can also suggest methodological choices, e.g., the family of citation method to select, depending on the objective, taxonomies of disciplines or finer level research fronts in a broad sense.

Shape/properties of clusters. Ex-post supervision of clusters (built by unsupervised methods) is a critical stage of studies. Discussion on the cluster aggregate features, or sampled articles, is much easier if clusters are reasonably homogenous. Therefore ability to recover clusters of any shape (elongated, non-convex...) which is essential in other contexts (say image-analysis) may not be desirable in bibliometric mapping. A few strongly linked compact clusters is easier to assess than the equivalent elongated class. The skewness of clusters distribution is another concern, especially in citation clustering, and the inflation of micro-clusters with poor connexions is unconvenient — an argument voiced in favor of direct citation approach for high-level taxonomies. From this point of view, the slight tendency of average linkage towards homogeneity and the tendency of K-means towards size-balance, giving moderately skewed distribution of cluster size, may be seen as "desirable biases" (see [146] in the context of community detection) with respect to further cluster supervision. As the benchmark exercise has shown, this does not prevent average linkage from recovering heterogeneous structures.

Soft vs. Hard Clusters. For reasons of convenience and computer efficiency, hard clustering is widespread but remains a violent approximation of the complexity and intrication of communities networks and semantic relations in scientific literature. Hard clustering is sometimes the first stage of a two-stage classification: co-citation analysis usually combines hard clustering for cores in the cited universe, and assignment of the citing literature tantamount to soft clustering of research fronts. Reciprocally, starting from hard bibliographic coupling clusters makes it possible to generate a soft image of cited clusters. The conditions of assignment parameters in the second stage determine the degree of overlap. This is true also for factor analyses more suitable for overlapping en-

tities, especially with oblique factors, i.e., principal axes of clusters upon which any entity, in or out, has a projection. The query expansion or bibliometric expansion practiced at the cluster level also builds soft clusters from an existing hard partition on the same data, therefore enhancing the recall at the cluster level. More generally, the wide development of probabilistic clustering is consistent with fuzzy approaches of assignment of particular articles/items.

Multi-level visualization of partitions is valuable for discussing topic or domain borders, especially when obtained from techniques which do not favour cluster homogeneity, or exploring strongly multidisciplinary phenomena. For example, assuming a strong proximity of two topics A and B, it is interesting to know whether this proximity is localized — say to sub-clusters A1 and B1 — or distributed. Local intense linkages may prefigure capture of a subcomponent or merge A1-B1. Such interpretation only makes sense with robust methodology.

In a cluster selection process for delineation, all things equal, soft or fuzzy clusters are allowed to extend towards shared areas, and then slanted towards recall at the cluster level. This applies to the boundary clusters, with an effect on domain's delineation. Bibliometric use of soft clustering remains however limited and does not usually depart from the holistic perspective (see "Semantics, Statistics, Informatics" below).

Robustness and evaluation issues. Robustness is an essential aspect of data analysis applied to bibliometrics. Sensitivity to data issues, to the type of network, to metrics and clustering algorithms, lead to rather different solutions. Ground truth or even gold standards are generally unavailable. In empirical studies, analysts have to get along both with biased representation of panels and divergences of techniques, as well as sensitivity to settings within one technique. We already mentioned general problems of bibliometric data, especially coverage. Within a given data corpus, the skewness of informetric distributions is a powerful foundation of robustness, but many sources of instability remain. The particular question of time robustness is sketched later. The particular question of time robustness is sketched later.

Sensitivity to the network weighting and metrics. For memory's sake, some prior transformation of bibliometric networks is practised to compensate across-domain differences, such as citing behavior. In such case, the value of linkages are weighted by a function of the number of inlinks of given groups (tantamount to classical cited-side normalization) or the number of outlinks. The latter is present both in influence measures (Pinski & Narin [44], revival in the last decade, e.g., [180]) and the limit case of citing-side normalization which presents original properties [181, 182]. Citing-side normalization of the citation network is a limit case (removing iteration) of Pinski & Narin influence weights [44]. It is strictly classification-free if the basic normalization unit is the paper or the journal [181]. It exhibits interesting properties for any basic unit making sense, e.g., domains: the dispersion of domains' impacts calculated this way with normalization at the domain level is a measure of interdisciplinarity

of science in a steady state system [183].

A major native characteristic of bibliometric networks is the skewness of node degree distribution and resulting polarisation: citations, Zipf-Mandelbrot words usage, Bradford concentration — in connexion with concentration generating models recalled above in social network theory. Concentration gives tremendous selective power and at the same time, calls for corrections in IR context. for information retrieval and usage, depending on the context. A vast choice of metrics or quasi-metrics (similarity indexes) is available, introducing weightings with some inverse function of frequency, especially useful in a mapping context. It is common knowledge that various similarity indexes produce contrasted perspectives. Co-word analysis pioneers, notably, compared the unweighted index (raw), the asymmetrical (inclusion) index, the partially weighted index (Jaccard, Ochiai among others), the strongly weighted index (p-index or affinity amenable to a similarity). After thresholding, the landscape of the transformed networks is quite different: the first two indexes tend to keep the frequent items as hubs, the last one highlights infrequent words and associations at some risk of overexposure of rare forms, amongst them typing errors.

Analogous normalizations, from the abundant repertoire of similarity indexes, are frequent for co-citation [184] and co-authorship analysis [185, 186]. Clustering algorithms build on the final network in various ways. Obviously, any delineation based on such weighted networks of structuring elements — where skew distribution is the rule — will be quite sensitive to methodology. In bibliometrics, the contrast is extreme between steep landscapes generated by raw measures, dominated by the centrality of hubs, and information-driven strongly corrected configurations, at the risk of instability and errors on very low frequencies. Intermediary options are often picked, for example Ochiai-Salton and Jaccard measure. Document coupling relations, similarly, depend on the normalization of terms frequency, typically inverse frequency weighting, Hellinger, etc. built-in or not in data analyses methods (tf-idf, χ^2 in Correspondence Analysis, etc.).

Asymmetrical relations. Specific to citations, a complete model of citation exchanges requires some native or constructed aggregation with relatively stable entities (authors, journals, pre-existing categories, etc.) in order to allow both in and out-linkages while document-level direct citation is unidirectional — with exceptions. Asymmetry at the journal level inspired the CHI classification of journals after their theoretical vs. applied orientation [44] on the hypothesis that applied science journals tend to import knowledge and export citations, and reciprocally for basic science journals. The same phenomenon appears at the field level (cell biology vs. medical research, for example).

The valuation of bilateral relations calls for methodological choices which can largely affect mapping and delineation. Take the simplest case where i and j denote two aggregates (journals, domains...) and assume the ij link normalized on the basis of the total outflow of i and the total inflow of j, and conversely for the ji link. Let us calculate the bilateral link between i and j

by the arithmetic mean, the geometric mean and the maximum of these two unidirectional normalized flows, a simplified variant of [87, 187] for the sake of the example. Should these valued networks be used for delineation purposes, they would tend to produce rather different results. The multiplicative indexes trivially penalize one-way relations typical of vertical channels, and tend to group entities with balanced relations, either particularly integrated channels or basic science fields with multi-disciplinarity relations, or else clients sharing methods or products. In contrast, the maximum index tends to retrieve vertical channels (say cell biology–medical research) regardless of flows dissymmetry. Additive indexes stand in intermediary position, and appear as a middle-ground choice.

Semantics, Statistics, Informatics. Scientific domains at the meso-level represent a considerable amount of data, especially in longitudinal series. The computing requirements, even with sparse bibliometric matrixes, are high, driving towards clustering or spectral analysis algorithms with high efficiency. The trade-off between computer efficiency and semantic power is far from simple. Correspondence analysis [134] was amongst the first factor technique to exhibit some semantic power in textual applications, especially a robust capability to group quasi-synonyms with the distributional equivalence property. In its wake, post-factor analyses keep claiming some semantic power (see above topic modeling) and built-in mapping capability. In parallel, local similarity techniques associated with traditional or innovative clustering methods from network analysis privilege the native graph of proximity and elements/links groupings. In those approaches the duality [structuring elements × documents] needs assignment decisions (e.g., research front assigned to co-cited core) with a semantic dissymmetry as to the internal scrutiny of clusters: while the detailed map of structuring elements is appealing for cluster evaluation (cited cores; within cluster word-map), the document coupling map, internal to a cluster, is hardly interpretable alone as stressed before.

Now, if word-maps present high potential for sociological interpretation, mere lexical associations remain semantically shallow with regard to truly semantic analyses. A common limitation to all these methods is the "bag of words" overlooking the rank of words and the structure of statements — downside partly alleviated by multi-terms treatment (noun phrases). Citations present a fuzzier relation to semantics (Sect. 3.4) but co-citation cores are nevertheless understandable for experts. Labels or lists of descriptors directly issued from co-citation or co-word cores, for example a ranked list of specific terms, or indirectly rebuilt from clusters obtained by coupling, are common but limited auxiliaries for evaluating clusters. Cards might be reshuffled with new competitors to LSA such as Neural Word Embeddings (Sect. 3.3.2). In addition to the similarity calculations in the word–context, useful for information retrieval, semantic calculations on word vectors are possible, allowing good performance in analogy tests (i.e., "Find X so as X is to A what B is to C") or inference operations on these vectors, such as `"king" − "man" + "woman" → "queen"`.

This gain in semantic precision suggests that, applied to scientific corpora — now increasingly available in full text — it could allow in the future for an analyst to select the semantic dimensions relevant for delineating scientific fields and constitute crisp or overlapping groups of articles (or parts of these) in this subspace.

A recurrent problem of more traditional bibliometric representations, a counterpart of statistical simplicity and computer efficiency, is the holistic character of linkages, especially if combined with hard clustering. In document coupling techniques, either word-based or citation-based, the standard linkage measure is the weighted and normalized number of words shared. In lexical coupling, an implicit hypothesis is that the (weighted-normalized) number of shared tokens reflects the dominant semantic dimensions of the paper. For example, if very few words or references refer to methodology, this dimension will contribute less, all things equal, to the shaping of bibliometric similarity, which can be misleading. In the opposite case, if methodology markers are prevailing, a transdisciplinary corpus will tend to be split between hard science literature and soft science literature on the domain, whereas mixed clusters would probably reflect the domain structure in a better way. Were the linkage between two clusters needing explanation, this should be inferred from the features and given the titles of the two clusters, unless the technique includes indicators of contribution. In clusters of structuring elements (word graphs, co-citation cores) the relations are interpretable when zooming on the fine-grain networks of words or cited articles, but without semantic characterization.

In delineation context, a minimum of semantic break up would make the scrutiny of the border region easier and faster. It could especially orient discussions on preferential extensions of a core zone towards neighbour clusters with shared methodology but new objects, shared object with new methods, etc. Ad hoc simple characterization of vocabulary has been successfully applied for other purposes, e.g., the level of application of biomedical research journals (see [188]). However, manual semantic tagging is quite intensive and field-specific. At the document level, many natural sciences articles can be labelled with simple semantic combinations. In computational linguistics, many works since Teufel et al. [189] (argumentative zoning) address this issue of categorization of scientific discourses and automatic annotation, applicable for example to the summarization of scientific texts. Several proposals on categorization of arguments have been made, many of them at the experimental stage. Liakata et al. [190] developed and automatized the Core Scientific Concept (CoreSC) categorization whose first layer distinguishes 11 categories: objective (Hypothesis, Goal, Motivation, Object), approach (Method, Model, Experiment) and outcome (Observation, Result, Conclusion). This line of research is extremely promising for bibliometric studies, especially domain delineation, but remains for the time being limited to small universes. In the meantime, oversimplified semantic indexing would help a lot in qualifying interdocuments or interclusters relations. Fig. 3.4 shows a fictitious configuration where documents are naïvely described by semantic triplets with various degrees of kinship. The graph display could be replaced by a superimposition of three partitions, each one upon

a different semantic dimension.

More intensive semantic mapping relies on sophisticated ontologies, knowledge models, semantic networks. If such resources have not been established beforehand and published, bibliometric studies cannot generally afford such heavy developments, however see [191].

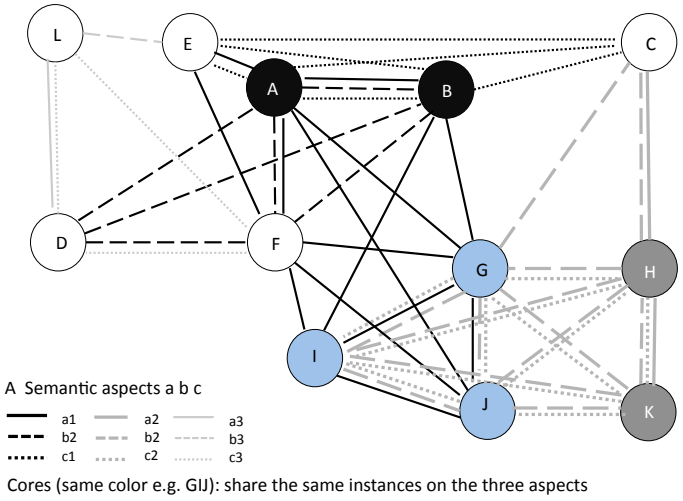
Directly opposed to semantic approaches are non-feature methods from computer science, which ignore the substantive representations and even more so the semantic content. In various IR/bibliometric applications (disambiguation of authors and affiliations, proximity of documents, detection of plagiarism) similarity between texts may be calculated on the basis of character n-grams [192] rather than "feature" word n-grams which is somewhat standard. The link to the minimal unit with semantic load, the word, is lost (almost completely for low values of n). Usual metrics can be applied to n-grams. A more radical way using the bit sequence representation with further compression, is the basis of measures like Normal Compression Distance (NCD in [193]). NCD is a dissimilarity measure which is an approximation of the general Kolmogorov information distance [194, 195], parametrized by the compression algorithm. A "normal" compressor should satisfy four properties: idempotence, monotonicity, symmetry and distributivity. From the linguistic point of view the compression method is a black box. It nevertheless exhibits rather good performances for calculating texts similarity with a most indirect semantic power of forms unification. The Normalized Google Distance (NGD in [196]) is the transposition to Google searches, at the word level, of the NCD, keeping the "feature" characteristics of the co-word analysis and its semantic power. Its native application builds on lexical associations from millions of users.

Table 3.3 summarizes the degree of semantic ambition in the case of lexical approaches — transposable to citation attributes.

Dynamic Clustering. The delineation process has to face changes in the configuration of networks [124], affecting the value of a delineation solution at a particular moment. Dynamic clustering is understood in two (related) acceptations.

A first point of view is the adaptation of algorithms — and computer resources — to processing massive data streams, typically texts, an example today, online social networks. The initial k-means algorithm of MacQueen [197] was already an "online" incremental one, generating a cluster structure in one pass over the dataset — the usual iterative version, for itself converging to a solution independent from the presentation order of the data vectors, is due to Forgy [198]. Dynamic text streams mining is a growing topic in the research communities of machine learning and Big Data mining. Changes in the cluster structure may reflect algorithmic artefacts as well as real phenomenon, hence ideal methodological characteristics are non-local optima seeking and independence from data ordering. An example of incremental hierarchical clustering method for texts is [199], and a frequent itemsets dynamical clustering is [200].

A second point of view focuses on domains/topics picture and their de-



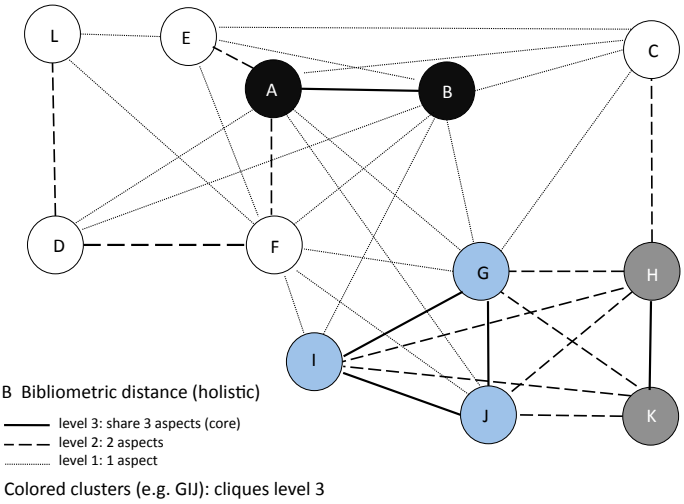


Figure 3.4: Semantic and bibliometric linkages. This figure sketches bibliometric holistic distance vs. decomposition into semantic links, with the (heroic) hypothesis of tagging with only three criteria, e.g., a = theory-hypothesis, b = experimental method c = observation-test. For example a1, a2, a3 figure denote different hypotheses. The second panel represents three kinds of semantic relations. An article is described by a triplet a, b, c. For example the documents G, I and J are described by the same triplet {a1, b2, c2}. Documents G and I, for example, are connected by three links. The second panel aggregates information in a single type of linkages with varying degree of intensity. Here the bibliometric linkage is assumed proportional to the number of shared semantic instances, which is of course arbitrary. In the real bibliometric world, the lexical coupling linkage heavily depends on the most developed aspect(s).

Table 3.3: Semantic thr. featureless

structural items metrics	document metrics (required in delineation task)	category	semantic interpretation vs. black-box
semantic network	indirect through indexing/ assignment to word structures	feature	strong, feature
word profile/ co-word	indirect through indexing/ assignment to word clusters	feature	light, direct
document profile/ coupling	direct: lexical coupling	feature	indirect, through indexing/ assignment/ labeling
–	direct: char n-gram proximity	featureless	black-box*
–	compression distance	featureless and global	black-box*

* clusters of documents based on non-feature proximity
can be interpreted by going back to substantive elements, e.g., their word profile.

scription over time, through clusters time-series, including the issue of time-robustness in one-shot pictures. Again the distinction between clustering/mapping on "structuring elements" (e.g., co-cited articles or lexical relations) and "direct clustering of literature" (e.g., bibliographic or lexical coupling), matters. The first family offers solutions with some durability. The repertoire of words gradually evolves. Change of the intellectual repertoire of cited literature, subject to ageing process, is usually faster but, except in emergence or revolutionary fields and in intrinsically rapid ones (e.g., computer science), it respects a mix of new and old literature. This gives some clue of robustness, in the short term, to the cluster solutions. By and large, in slow evolution processes, information cores are more persistent than peripheries. In one-shot clustering, working on pluri-annual window data reinforce the robustness of the breakdown and permit the cross characterization of novelty (median of the co-cited core) and internal growth in the span of the window (average date of front) [176]. Characterizing fine-granularity hot spots in the network, such as local preferential attachment processes, may help to spot promising weak signals. Taxonomic applications of direct citation linkages might still benefit more from long time-windows settings. This would sketch, as noted before, a possible trend towards division of tasks between direct citation, co-citation and bibliographic coupling in function of targeted granularity and immediacy of results.

By construction, direct clustering of documents over a time-period (say the year) favors immediacy, but is not prolongable without a detour by the structuring elements and derived cluster labels. Another way consists in picking a coarser granularity, especially the journal level, at the expense of a heavy loss of precision. Short-time changes may be addressed by projecting a solution for a period on the reference solution of another period, a classical process in factor analysis applicable to other methods, an early example within bibliometrics is found in Noyons & van Raan [32].

A delineation process of any kind may be run on successive slices of time (e.g., [201]) of different lengths, with or without rolling averaging filters. A dynamic variant of LDA is [202], in which the word distributions of each topic varies in each time slice, where the number of clusters is fixed. Interesting historiographic insights accounting for clusters demography (emergence, death, splitting, merging...) are exhibited by longitudinal chaining of clusters, known since ISI's Atlas of Science, see [203, 204, 205, 206]. The latter work is based on lexical series. The predictive value of such series, along with life-cycle models, remains a quite difficult issue.

Last but not least, the rendering of change is closely linked to dynamic models of science where structure emerges from local properties, for example in the preferential attachment model. In this view, over time, breakthroughs (scientific or technological) shape the citation profiles of followers, a common mechanism in (co-)citation bibliometrics. Local accretions around hot papers are amongst the signs of emergence. The symmetrical question over whether the referencing (or lexical) profile of papers has some predictive value, remains open. This connects to the controversies about interdisciplinary distal transfers in the discovery process, quoted above, which echo the combinatory nature of

invention and innovation stressed by Schumpeter. The intuitive but bold hypothesis stating that the more distant the knowledge transfer, the more radical the discovery or invention is, nevertheless, tricky to test (definition of scientific or technological distance from models A or B–C, scale issues). Attempts to characterize scientific breakthrough and radical inventions, with an *ex ante* notion, are found for example in [207], using both citations and patent classification; [208], using changes in forwards and backwards citation profiles; [209], using citation contexts of outstanding discoveries.

3.3.3 Conclusion

By and large, bibliometric mapping provides landscapes with aggregate groups (clusters; local factors, etc.) likely to be assessed, and implementation of multistep and cross points of views help to distinguish cores and border regions, the latter calling for cluster evaluation, see Sect. 3.5.2. No mapping method is superior on all criteria and many factors are at play: the bulk of data, the type of network, the nature of the problem and the ergonomics of outcomes for an easy supervision. IR search remains an alternative or a valuable complement to mapping. The next section zooms in on hybrid techniques.

3.4 Multiple Networks and Hybridization

This section addresses the multi-network approaches. We shall especially develop the combination of textual and citation networks but most types of bibliometric (and altmetric) networks can naturally contribute where appropriate. The forms of hybridization encompass a wide scope from fully integrated approaches to parallel schemes aiming at comparison and eventual combination, with intermediate sequential schemes.

3.4.1 Multiple Networks

A given document may be accessed by search strategies pointing at all searchable fields of data or metadata. Modern IR, going beyond the direct query-document similarity, integrates, with the cluster hypothesis and later the cognitive model, the documents' multiple spaces and networks, including citations and collaborations. Bridges between lexical and citation universes were built, especially for labeling purposes (e.g., keyword-plus [210]).

Likewise, major streams of study in the sociology of science have coined general theories accounting for the various manifestations of scientists behavior in communities: communication, collaboration, publication, rhetoric, citation, evaluation. The networks of science, although diverse, originate in the same ground. As a result, many classes of bibliometric questions (topic identification, characterization of emergence, static and dynamic mapping, diffusion processes, knowledge flows in science and more generally in the science-technology-innovation system) can be answered by working on different networks, with

respect to their specificity. The multi-network approach to bibliometrics, both in terms of comparison and complementarity, appears as a natural mode of thought.

With the coming of age of data representation models such as entity-relationship for Relational DataBase Management Systems (RDBMS) implementation and of network analysis methods, IR scholars and bibliometricians in the early 90s found flexible tools for easy handling different dimensions of publication data. In the last decades, the culture of data-mining encouraged mixes between several networks for pragmatic purposes [211]. We recall the key role of authors networks (Sect. 3.4.2) before focusing on text and citation networks (Sect. 3.4.3) and finally their hybridization (Sect. 3.4.4).

3.4.2 Actors Networks

The first analyses of scientific communities in the seventies lead to some disappointing results as to the unambiguous assignment of particular scientists to a particular group. In a short history of domain delineation Gläser et al. [26] recall among others Mulkay et al. 1975 work [9] and Verspagen & Werker findings [212]. The archetype is the co-authorship graph. Price & deBeaver [18], deBeaver & Rosen [213], Luukkonen et al. [214], Kretschmer [215], Katz & Martin [216] laid the first layers of collaboration studies in connection with invisible colleges. Authors-based models of science are amongst the central topics in science studies and bibliometrics. Studies on scientific collaboration are out of our scope here, let us just recall the macro-level studies of the determinants of cooperation in the wake of Luukkonen et al. [185], geographic proximity [217, 218], cultural links [186], individual/collective behavior [219]. Those studies emphasize the importance of metrics and normalization in the interpretation. At the micro-level, proposals for mechanisms explaining the structure and dynamics of social networks were recalled in Sect. 3.3.

Actors' networks present a major theoretical interest: they stand at the crossroads of actual social networks' mathematical modeling and sociology of research, and bridge invisible colleges with cognitive structures [220]. They also show some drawbacks, echoing the scholars' disappointment noted above. Communities detection in practice faces the issue of names unification. The problem has been for a long time terribly cost and time consuming for data producers and bibliometricians, at both the institutional level and the author level, as stressed again in the Name Game project APE-INV (`http://www.academicpatenting.eu`), e.g., [221]. Great progress is ongoing due to ORCID (with the unique identifier of researchers), ISNI, GRID initiatives among others.

Another issue, especially for small topics detection, is the width of the competence spectrum of productive authors likely to produce some noise, but this shortcoming is alleviated at the level of large domains. In this case perhaps, community detection (in a narrow sense) has arguments to compete with citation or lexical clustering. However, in most practical studies multiscale vision is required: not only does the target domain matters, but also subdomains. At this scale, the polyvalence of authors limits precision. The problem may be

reduced by time-restriction filters, link-level technique, external information or hybridization with citation or word information. Similar issues appear in "author co-citation" vs. "article co-citation" [164, 222]. Author co-citation opened insights in the study of invisible colleges, with connection to researchers' sociology. Topics mapping as such is better addressed by document-level co-citation.

The interplay of co-authorship, citation and linguistic networks as a mirror of socio-cognitive activity is increasingly gaining attention: relations between contents and actors' positions [223, 224], between citations and co-authorship, and any or both of these with texts [220]. Is the multiple approach-a step towards more powerful models of authors and community behavior, able to unify the diverse representations? This unification would spread benefits over bibliometric analysis, including delineation tasks. Non-feature methods have not awaited for unification (see below) to mix up all types of information, but they sacrifice the substantive depth of analysis.

However, the quest for unification might be hindered by the specific features of every bibliometric network. Changing the type and parameters of the network is like observing the universe in various wavelengths. The most dense objects produce various forms of energy and tend to be retrieved albeit with diverse volume and appearance. Less dense objects like clouds of various composition can be seen only in specific parts of spectrum. Likewise, we may conjecture that dense and isolated objects will be retrieved upon any network fit for precise analysis (e.g., [113]), especially words and citations and perhaps co-authorship clusters. Sociological investigation is expected to confirm such configurations as bounded invisible colleges. In less dense and more connected areas, each network is likely to produce non-superimposable images, with different sensibilities. The convergences suggest strong forms with easy socio-cognitive interpretation, while the divergences ask for careful tests and investigation. The sociology of translation associated less dense areas to emergence or ultimate evaporation phases.

3.4.3 Citations and Words

Lexical and citation characterization classically used in bibliometrics are appropriate for themes clustering and mapping at various scales, on the basis of the toolbox sketched in Sect. 3.3.

A few Analogies and Differences

General. One difference naturally lies in the nature of the original relation: direct attributes for linguistic elements, reflexive inter-articles for citation, with several consequences. Firstly, the granularity: words are an ultimate attribute (in classical "feature" methods) whereas cites target the full article semantic aggregate. Then, the linguistic content of citations is not explicit, and requires a statistical detour via the text fields and the data model, to emerge (automatic labelling of clusters with their specific vocabulary, citation contexts). Secondly, the time relation, not explicit in lexical relations, directly appears in the cita-

tion link, both cited and citing article being dated. Bibliometrics makes a large use of this diachronic relation in immediacy-ageing studies. In contrast, the word content of an article is readily legible, but deprived of temporal information beyond the article date of submission/publication. Going further requires statistical studies to date the word in terms of chronological profile of use. Longitudinal studies on words have to rely on time statistics of use, typically with the assumption of achronicity: constant meaning over time. This is a bold statement in some cases. Beyond classical dating of word or word linkages after their usage, determined by the obsolescence of topics, natural language analysis paved the way for analyses of word transformations in a scientific context [225].

With respect to these constraints, a large class of bibliometric, IR or altmetrics issues can be addressed by the lexical way or the (generalized) citation way with the exception of specific direct chaining [162]. Symmetrized relations (cocitation, coupling) mitigate the diachronicity, albeit underlying time features can be invoked if required. The reformulation of the dynamic chaining research fronts (e.g., [205]) is emulated by word-based clusters [202, 206]. Only the former directly contains citing-cited information for immediacy characterization.

Due to limitations (indexer effect) and lack of reactivity of controlled language, modern bibliometrics moved gradually towards natural language, building on the increasing availability of full text resources and lexical treatment. In spite of progress in computational linguistics, the NLP remains tricky, a counterpart of language richness and versatility. Polysemy, metonymy, synonymy, figures of speech, metaphors, acronyms and disciplinary jargon are well-known linguistic traps of linguistic difficulties that users, bibliometricians, retrieval specialists have to cope with. Unification (stemming and lemmatization, synonymy detection) also benefits from clustering techniques. Unsupervised homonymy tracking is a more challenging problem, since bridges in word clusters may be rooted in concept transfers or polysemy or else simple homonymy. This issue is somewhat alleviated in small (narrow context) studies. If elaborate ontology or semantic networks are seldom off the shelf, useful tools for term extraction, parsing, co-word exploration are available. Stemmers (with the Porter's stemmer milestone [226]) or, a step further, lemmatizers are efficient with some risk in precision. New massive techniques, such as above-mentioned deep learning based or Neural Networks or targeted methods such as Neural Word Embeddings, might bypass or alleviate costly preprocessing. Constraints of bibliometric studies dealing with large data universes are usually incompatible with refined semantic treatments, but the supply of large-scale statistical semantics resources might spare costly ad hoc developments. We mentioned (Sect. 3.2.4) a possible revival of controlled vocabulary supported by bibliometric treatments.

Statistical Background. The common feature is the skewness of frequency distribution found, among other disciplines, in information processes (Bradford-Lotka-Zipf trilogy, see [227]). The classical model to fit word distributions is the hyperbolic Zipf-Mandelbrot model. Other Paretian distributions are also used for citation frequency analogous to node degrees in the native oriented

graph of citations. Similar skewed distributions are found in authors' collaboration graphs, with a distinction between scale-free distributions and small-world distributions (see Sect. 3.3). The parameters of citation distributions are modulated by the citation-windows, the parameters of word distribution modulated by the type of lexical sources (title, abstract, full text...), the type of lexical unit picked, the language, the richness of vocabulary.

Comparing the distributions of citations and words on the same corpus, some authors found that the latter appears more concentrated and less 'complex' [33], thus less favorable in principle to precision — without forgetting the different granularity. Frequency weighting of linkages of the native word or citations networks, or similarity indexes with various types and degrees of normalization, may be implemented for retrieval or mapping purposes, for favoring information-rich elements in low and/or medium frequency. The precision of citation approaches was underlined in comparative retrieval tests, and especially the interest of cross retrieval [228, 229]. As to co-occurrences, co-word matrices tend to be less sparse but noisier than co-citations relations.

For the delineation work, the distribution of words or citations designs the background, with implications for interpretation, but what directly matters is the arrangement of documents after their texts or their bibliography. For this purpose, the typical ways are the direct profile proximity on either type of structuring elements, words or references ("coupling" rationale or profile metrics in vector space), or the secondary assignment on prior classes of structuring elements such as co-word, co-citation or corresponding profiles. The distribution of node degrees in bibliographic coupling tends to be less skewed than in the original citation graph. Again normalization of distances or similarity by some function of inverse frequency can reduce the unevenness. The recall advantage of word-based techniques suggested to use them in the large-scale mapping of clusters defined, beforehand, by citations [91]. There is some evidence in the same direction for patent-publication relations. Composite word-citation metrics are addressed in Sect 3.4.4. Technicalities involved in term unification are also different. As information tokens, references are less difficult to match than natural language elements. Keys on cited references reveal effectivity and improve with standardization of entries, with residual difficulties in particular cases like citation analysis of patents towards science.

Sociological Background. The textual contents of an article and its bibliography are both the results of authors' choice in their community context. Both involve an intricate mix of scientific and social aspects: words and cited references are community markers and reflect the sociability of invisible colleges. A large body of literature (see the review [230]) has been devoted to citation behavior, including Cronin's classic work [231]. Whatever their determinants can be, Merton's rewards, Small's symbolic beacons or concept symbols [232], Gilbert's persuasion tools [223] or Latourian interests, the references mainly point towards the thematic groups where founding fathers, gatekeepers and potential partners are found, which matters in science mapping. On the textual side,

rhetoric and jargon expressing community habits, in addition to general words voicing interests, rejoin focused scientific terms — especially specific multiterms with medium frequency — to define topics. A substantial amount of convergence between texts and citations is therefore expected when the delineation of topics and communities are at stake. Some degree of parallelism may be found between relatively high frequency expressions (after filtering of stop-words) and highly cited articles in generic knowledge and multidisciplinary linkages. The measured convergence depends on the information unit and is likely to increase with small lexical units of citation contexts (see below).

However, the question arose of which network is the more appropriate for describing science, at a time (the eighties) where citation evaluation, indexing and mapping were gaining interest. The social constructivist stream and the Actor Network Theory mentioned above (Sect. 3.2) favoured the co-word networks [166] against citations to represent knowledge on a background of actor's interests. Texts appeared abler to depict more completely "science in action" [233] especially in controversial areas where social and cognitive aspects are inseparable, while citations were supposed confined in the capture of "cold science" with delays and incompleteness. The delay argument alone is less convincing for bibliographic coupling. Typical co-citation "research fronts" rely on a high-pass filter on citation or co-citation scores, favoring old articles, to reduce the data volume. Bibliographic coupling often works on the whole reference lists, letting recent and less cited references play. A residual effect of the publication cycle of the citing side nevertheless subsists. Similar delays may also occur in the use of new words or expressions qualifying a scientific technique.

In its very realm, academic science, citation analysis encountered lasting problems in quite a few disciplines, especially in a fraction of SSH, because of citation sparsity, uncomplete processes of internationalization and lack of coverage in databases. This argument is somewhat weakened nowadays because of data source progress and changing behavior of scholars confronted with the science globalization and bibliometric evaluation. Citation analysis proved an appealing tool, including for the borderlines of standard literature, for example transfer documents (guidelines and even magazines and newspapers) explored in biomedicine by "translational research" for improving health systems services [234]. See also the EUSTM website at `https://eutranslationalmedicine.org`. As to the coverage of technology, the transposition of citation analysis to patents was revealed to be rather successful [235] competing with lexical approaches [236]. Of most importance perhaps, the Internet produces linkages with an exploitable analogy with citations, as the Google search engine has demonstrated in the wake of Pinski & Narin's influence weights.

Citations are not without their shortcomings, stressed in voluminous literature from various horizons (see the extensive Bornmann & Daniel's aforementioned review [230]; for the defense, mostly, see [51]). For the reason stated above, biases of citations are somewhat less severe in mapping applications than in citation evaluation (impact, composite indexes) which concentrate controversies. Latourian citations or rare negative citations do not add much noise to

co-citation topics. Other down-sides are more serious. The bandwagon effect in citation behavior tends to create spurious cliques in native co-citation networks, possibly hindering the discriminating power of citation relations. The inflation of the number of references in authors' practice, which is a long-term trend [237] also brings noise to conventional citation clustering. The disciplinary insertion affects the number of references ("propensity to cite") justifying citing-side normalization approaches mentioned above.

Albeit language-dependent, textual analysis is media-free, which is valuable in fields where academic sources with standard citation behavior are not sufficient. Topics peripheral to the academic mainstream, or demanding a mix of heterogeneous data may be confined to text-based delineation.

In cases where no differential data coverage issue is faced, differences may arise between these expressions of scientists' behavior, resulting in alternative breakdowns into topics, independently from statistical properties. The expectation is that citations, albeit in a blurred and biased way, are more capable to tracking the intellectual inheritance. A single difference in the semantic mix, for example different methodology on the same category of problem, will probably better discriminate amongst micro-communities than lexical analysis, at least as long as those micro-communities do not secrete specific terminology.

Let us turn towards limit cases, special forms of particularism, especially perhaps in SSH where intellectual traditions resist globalization. Words as well as citations would distinguish between schools of thought with opposing theories, strong community preference and distinct jargon: say in postwar period marginalist vs. marxist economists. In contrast, if the linguistic repertoire is shared by the two communities while they diverge in the intellectual base, the outcomes of the two approaches will be different. The reverse can be true, with a common recognition of the intellectual base but divergent traditions in terminology, perhaps again for reasons of national tradition. Such configurations, relatively rare, limit the generality of the conjecture stated above about local convergence of bibliometric networks in zones with high-gradient borders. Most of the time, a set of clustered papers belonging to a strong overlap of a word-based cluster and a citation-based may be considered as a strong form, in a rationale already present in the first comparisons by McCain [228] on term vs. citation indexing. The cognitive overlaps between information types was a keypoint in Ingwersen's model mentioned above.

Empirical Comparisons

Cross-check of clusters contents is run-of-the-mill operation. For example, the enhancement of co-citation coverage by two-step expansion could be controlled by lexical means [127]. A few specific comparisons of the two mapping approaches on the same data are found in literature. The scale is therefore different (sub-areas rather than a large domain) but the method can be applied to an overset expected to contain the targeted domain, as seen before. In an extensive study of a few promising fields in the 2000s, using bibliometric mapping, Noyons et al. [238, 239] warned about the difference of concepts: publications

and keywords and concluded to "totally different structures."

An opposite conclusion was reached by Zitt et al. [240] on nanosciences and Laurens et al. [241] on genomics, previously delineated as a whole by a hybrid sequence method. They implemented a more direct comparison scheme on clusters respectively from bibliographic coupling and lexical coupling (natural language, titles-abstracts), using the same Axial k-Means method (AKM). Cross-tabulate cluster overlaps (see also [242, 243]) were reordered, giving a quasi-landscape with a heavy and narrow diagonal load (Fig. 3.5). This gives evidence of a fairly good convergence of lexical and citation solutions, also confirmed by direct indicators.

On their high-level maps, Klavans & Boyack [244] and Leydesdorff & Rafols [245] also observe a reasonable degree of convergence. More general comparisons of mapping methods including textual are found in [173, 246, 247]. A recent exercise of mapping comparing clusters methods is reported by Velden et al. [248]. Most experiments however lack a ground truth reference, and techniques presented as gold standards are disputable.

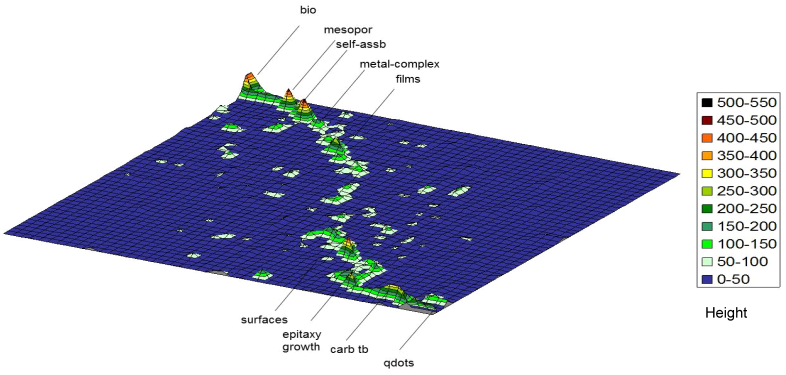
More generally, suppose we built clusters of documents from several origin: lexical coupling, bibliographic coupling, fronts from co-citation, author coupling, etc. Those various cluster solutions may be individually mapped. They can also be simultaneously represented using normalized overlaps between w-clusters, c-clusters, a-clusters, with an appropriate metrics. Profiles distance may be required to overcome the zero overlap between hard clusters of the same family say w-clusters. Resulting matrices are still quite small and amenable to MDS display.

The fact that the agreement between citation and lexical approaches is good but not complete brings one more argument in favor of complementarity. One thing to keep in mind: due to imperfect optimization of reordering and choice of the article rather than sentences or narrow contexts as the lexical unit, the global convergence tends to be under-estimated.

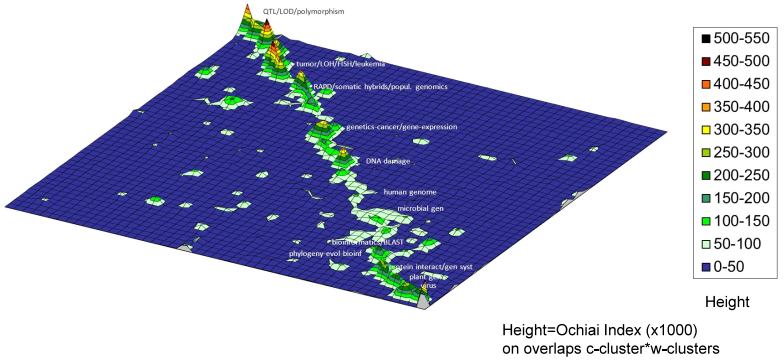
Complementarity

Complementarity, rather than competition, already inspired the "citations in context" researches, initiated in co-citation studies (e.g., [249, 250]) which are a natural space to connect referencing, intellectual base and linguistic aspects. In a step further than linguistic labeling entities in (co)citation analysis, the studies of "citation in context" range from simple context visualization in citation engines to investigations in the dynamics of science. They tend to reinvest research in action, associating language and communities' life. The linguistic and semantic analysis of citations contexts contribute to topics such as the citations types or motives [251], the classification and cross analysis of the contents of the citing or the cited documents [252], the fine-grain relation of citation contexts and abstracts terms [253], the exploration of new dimensions of scientific texts [254]. Some of these advances influence citation techniques in return. An example is the improvement of co-citation accuracy [255, 256].

As a result of multi-network or poly-representation hypotheses, some issues



(A)



(B)

Figure 3.5: Achipelago display: Nanosciences (A) and Genomics (B). Data: Reordered cross-tabulate matrix of Axial K-Means clusters respectively from bibliographic and lexical coupling 50×50). Relative overlap (z-axis) measured by Ochiai Index. Reordering: ranks on 1-dim MDS, making it apparent the diagonal accumulation showing the visual convergence between the two breakdowns. The line is sinuous because of discrepancies between c-clusters vs. w-clusters size-distribution. The visual rendering suggests superclusters at a larger scale. In the nano figure, the area of nanotubes as a whole is retrieved by both methods, but with two different breakdowns and more discriminative power on the citation side. Source: [113].

typical of one representation can receive a solution from the other. Convergence at the local level also creates spaces for complementarity: synonyms of any kind, for example, tend to be retrieved in the same citation-based clusters. Citation techniques escape linguistic polysemy and the reverse is true, but "citation homonymy" often due to matching keys, is a less important risk.

Finally, textual information preserves its advantages of availability, intuitiveness, and interpretation, with easy transposition to concepts and topics. A major shortcoming is the complexity and ambiguity of natural language, resulting in poor precision in case of unsupervised protocols. In spite of the composite unit handled (the full article rather than the narrow concept), citations are appealing for tracking intellectual influences and often less noisy, at the expense of lower recall in weak signal configurations.

The capability of pure lexical approaches to emulate citation-based or hybrid approaches in challenging topics such as the aforementioned description/anticipation of early stages of domain emergence, remains a challenge.

3.4.4 Hybridization Modes

Looking for optimal exploitation of these contrasting properties is the quest of hybrid techniques, in line with pragmatic mixes of dimensions in IR-type delineation for bibliometric purposes. The same pragmatism inspired mixed information classification of web sources [257]. The detail of the more sophisticated techniques are not on the table: millions of Google users benefit from hybrid IR processes every day, but in spite of expansive literature devoted to the PageRank algorithm itself starting with [258] and published works on lexical/semantic processing [196], the detailed combination of multi-network operations in the search engine is not documented. We will limit ourselves here to quite basic combinations, readily available in bibliometric literature.

The scope of hybridization is quite large: words and citations, on which we focus, may be taken either as variants of information tokens likely to be indistinctly treated under certain conditions, in a typical informetric posture; or seen as elements of quite different relations with their own fundamental properties and interpretation, suggesting to use them in sequential or parallel protocols. Parallel exploitation, particularly, is "sociology-compatible" allowing for separate interpretations and comparison before final combination if necessary

Full Hybrid

The structuring/clustering of fields using a common metrics mixing citation and term distances at the finer grain level, from the start, is a promising path [259, 260]. Boyack & Klavans [128] on a large dataset, observed that even a "hybrid naïve" coupling outperformed pure bibliographic coupling. Statistical differences between word and citation distribution can be reduced through a normalization of the similarity measures with different distributions ([261] later simplification in [262]) achieving a full and flexible integration. Koopman et

al. [263] established cluster similarities using a combination of tokens, for comparing clustering solutions based on direct vocabulary and indirect vocabulary associated with authors, journals, citation, etc.

Those developments remain in the framework of "feature methods" keeping the substance of information elements, words and citations. In Sect. 3.3 we mentioned purely computational methods (character n-grams on text flow, compression) for calculating generalized text distances regardless to linguistic features. An option is to stay within the textual domain (full text, abstract, title...) or to enlarge to the full article including authors, affiliations, list of references, etc. We get a massive and blind form of hybridization, dissolving both terms and references in signals, ignoring all forms of normalization including for zones length (text vs. bibliography). Such black boxes are deprived of any semantic interpretation, but in our experience prove efficient for quick calculation of inter-document distances.

We have seen above (Section 3.3.2) that Deep Neural Networks have proven in many areas of supervised learning, including information retrieval, their ability to do without prior weighting of the variables. Their unsupervised variants, building upon their success in very constrained fields like the Go game, should be able to do the same from an informal collection of data — such as "full hybrid" data — and so an application to domains delineation might be to consider the last layers of a network collecting the many traces of scientific activity, whatever citations, texts, and so on in the wake of present limited attempts of hybridization. Research in unsupervised deep learning, though, is still at a preliminary stage [264]. There is no doubt, however, that in the next years progress — and controversy — are to be expected from deep learning entry into the competition. These processes, however, remain black boxes, with quite difficult interpretations. Perhaps high-level semantic categorization resulting from the careful interpretation of the last layers might allow experts to select a subset of explicit dimensions in order to take into account the users' expectations of a delineation process. Could this reconcile cognitive classification and institutional expectations, an issue mentioned above, is another question.

Sequential Hybrid: Citations → Terms

Sequential protocols of delineation may rely on more iterations, we limit ourselves here to point out the basic sequences. We mentioned above the tradition of completing citation objects by textual tagging. The question of the validity of co-citation research fronts (see Sect. 3.2.3) triggered further developments on retrieval and recall rate and the means to foster it, possibly with the help of texts. Braam et al. [265] developed a systematic complementation of co-citation clusters coverage by lexical means, a first operational example of hybrid delineation. The citation → text sequence keeps being explored for other purposes, especially in global science maps. Boyack & Klavans [91] use textual metrics for display of co-citation cluster relations at the large scale where citation signals are weak.

Sequential Hybrid: Terms → Citations

The perspective is reversed. The remote ancestor is a classical application of citation indexing, when title words or keywords+ were used to query a citation index for harvesting papers on a given (set of) topics. The rationale is simple: starting a multistep process with experts' help is easier with word queries. In a second step, the expansion is carried out on the citation network, where unsupervised or lightly supervised procedures are safer than on texts, with proper precautions. General conditions for citation analysis are required, especially not too scarce reference lists. There is some analogy with the "boomerang effect" on citations [266]. An example of protocol is the lex+cite process explored in Laurens et al. [241], especially for emerging or transverse domains, where classical methods tend to fall short.

Quite a few options exist for expansion. If the seed is considered globally, literature with references combinations present in the seed, but not in particular papers, is recalled. However, not specific cites should be ruled out, which may require information from the whole database; conversely, if only combinations at the paper level are allowed (strict bibliographic coupling), a typical literature is missed; cluster-level enrichment, if a previous breakdown into clusters is available, stands in the middle. Besides the recall-oriented aim, these hybrid protocols may also enhance precision by submitting the core itself to bibliographic coupling constraints. In the same line, an elaborate strategy starting with lexical queries and query expansion, completed by journal selection and ending by collecting citing papers, is proposed in [267].

Parallel Design

As described above in Sect. 3.4.3, parallel design allows for comparison especially when metrics and clustering methods are identical, so that the final outcomes can be compared by factor analyses, parallel clustering-mapping and reordered cross-tabulations. In parallel clustering, a similarity between clusters from different origins is defined after their degree of overlap, and then the inter-cluster matrix, of small size, is easily displayed using a MDS-type method. The cross-tabulation for example highlights strong relative overlaps with two strategies in addition to choosing either the c-cluster or the w-cluster on a topic: (a) precision-oriented: a heavy intersection between c-cluster and w-cluster suggests a strong form of topic, strategy possibly extended to superclusters (c) recall-oriented strategy, taking the union of c-cluster and w-cluster.

3.4.5 Conclusion

The various publication-linked networks, at least words and citations offer globally convergent views but not at the point that one can be happy with a single solution: sociology of citing, collaborating behavior and writing rhetorics keep some distance, and bibliometric protocols can choose to mix up all information tokens or to combine parallel approaches at a final stage. Comparison and

complementarity merit further endeavor. In practice, delineation cannot avoid supervision and actors' negotiation. Protocols of experts' guidance for evaluation purposes are desirable. Cross-validation of parallel processes, and even in some cases of sequential processes [241] may alleviate the burden of multistep external validation. There are strong indications that multi-network methods improve the recall and offer richer substance to expert/user discussions, but more benchmark studies against ground truth are needed.

3.5 Delineation Schemes and Conclusion

3.5.1 Delineation Schemes

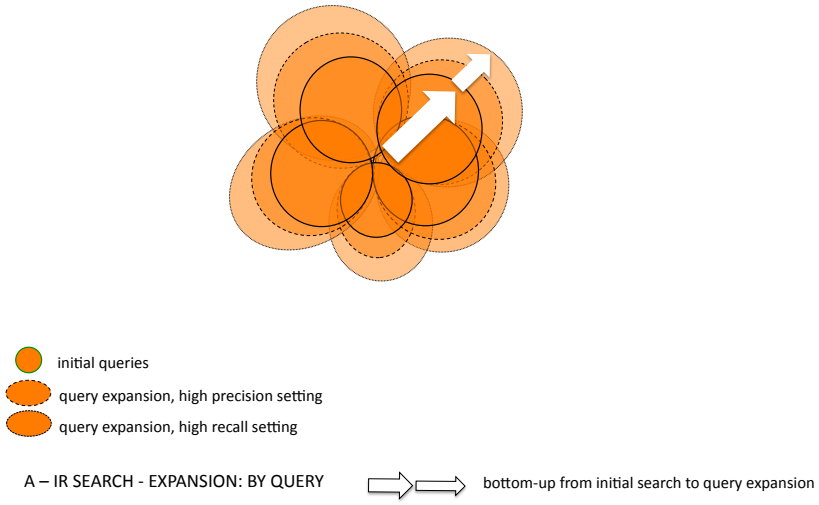
IR Search First

A scheme of bibliometric study asking for careful delineation may be as follows:

- For memory sake, selection of the expert/peers panel, matching the expected variety of the domain.
- Supervised IR search on specialized journals and specific vocabulary, aiming at precision, building up the core of the domain. Alternatively, use of cited cores at the article or author level. The granularity is, typically, the document level. In favourable cases, some partial query formulas are found in literature.
- Query expansion or bibliometric expansion with citations (the latter usually requiring lighter supervision). The query expansion is conducted globally or query by query. Optionally a round of data analysis/clustering can suggest rephrasing or complementing the queries (Fig. 3.6).
- Evaluation of outcomes especially on borderline. In multilevel processes, the border region typically stands between high-precision cores/seeds (or low-recall expanded set) and high-recall expanded set. Circles of expansion with expected relevance indexes (example in Sect. 3.4) enlighten decision, again optionally supported by thematic clustering/mapping.

Clustering/Mapping First

Regional overset maps are expected to contain all the target, and the decisions on border regions are typically made at the cluster level. Granularity obviously matters: we cannot expect that any high-level clustering of global or superlocal science will directly produce a class retrieving the target domain as a whole. A lower-level breakdown yielding fine-grain delineation of the frontier will be preferred, with a number of subareas large enough to match the diversity of the domain and eventually increase precision, but small enough to make cluster-level expertise feasible. Reasonably, the granularity picked fulfils two objectives, aiding the delineation and preparing the study of the domain's subareas.



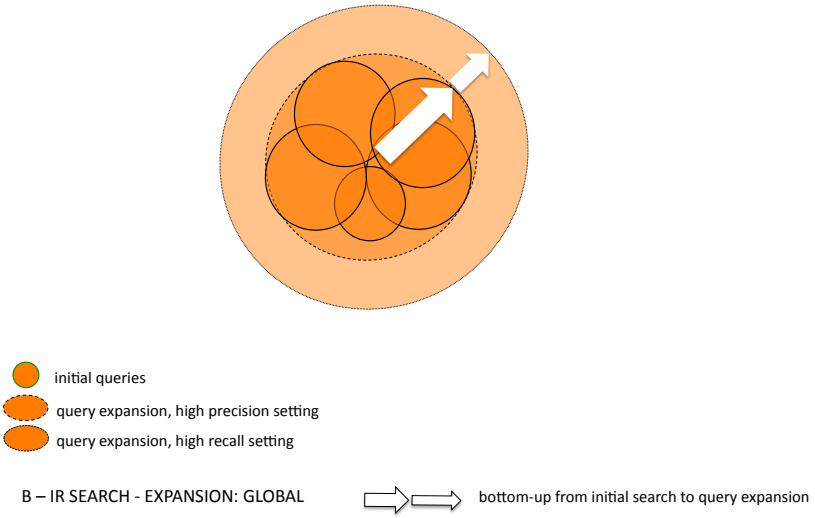
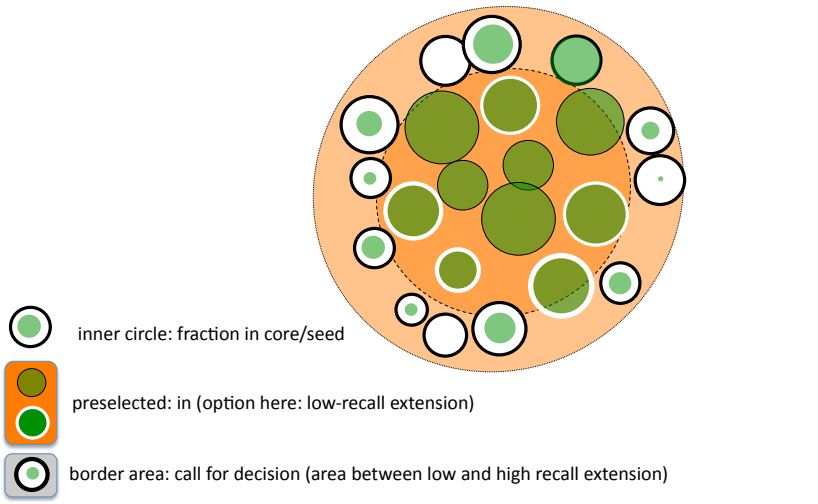


Figure 3.6: IR search and mapping approaches (1/3). (A) IR process: bottom-up queries and expansion of individual queries. Assumed at the paper-level. (B) variant of A: expansion based on the entire set (lexical or citation-based). The border area, to be discussed, is typically determined by the region between high-precision seed (or a low-recall expanded set) and a high-recall set. Circles of expansion in the border region, if indirect indicators or relevance are available, can drive the choice of delineation.



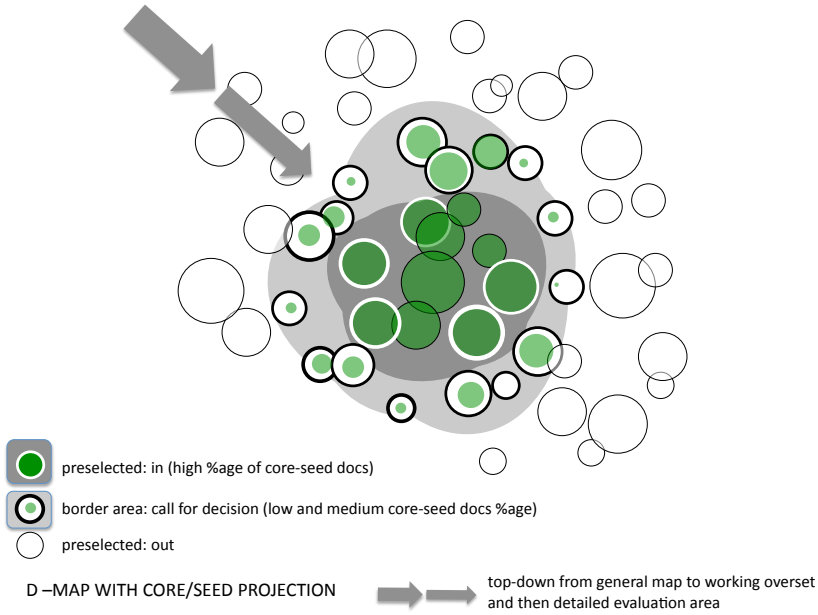


Figure 3.7: IR search and mapping approaches (2/3). (Optional C) local clustering from B data (alternatively: from A data). Clusters are helpful for discussion but the border region and the decision tools may exist in A or B stages. The map is local and in the general case is not superimposable to a fraction of the global map D. A discussion on global vs. local mapping is found in [179]. (D) global mapping/clustering: top-down from global or overset map to the target; detection of border area.

In the perspective of cluster evaluation procedure, possibly time-consuming and costly, it is recommended to rely on a lightly supervised preselection of the border region, located between the internal core, a priori deemed "in," and the external zone deemed "out." Depending on the clustering-mapping protocols chosen (see the sketch Fig. 3.7), various solutions can address this preselection, for example:

- Clustering with IR Search Projection. For this preselection, most helpful is the simultaneous representation of a global map (or at least of an overset-map) obtained on one criterion and cluster-level properties on another criterion. The projection of local features over a large context is often used: in two-steps protocols, seeds for example are projected on clusters in the expanded set [241] with the ratio of seed articles as indicator for delineation. Another combination: a global map conveys a particular vision depending on the network represented and the methodological choices made, and the hits of a IR search on a lexical marker (with a generous setting for recall) alerts on clusters of interest. In Fig. 3.2 for example the central communities might be considered as belonging to a core, whereas distant colonies, on the borders, require evaluation. Such cases illustrate the complementarity of IR and mapping techniques for avoiding silence both on weak and strong signals, as mentioned above. An alleviated process uses the projection of specialized journals literature onto a global map [177]. Such processes help pinpoint clusters forming the border region as "decision area" and/or suggest journals or groups of papers as candidates for extending a core. Clusters may also undergo a complementary stage of query expansion or bibliometric expansion, typically transforming — in a given universe — a hard partition into an overlapping structure. For the domain delineation, only the overlaps involving the border region will matter for the final outcome.

- Crossing methods. An alternative is the crossing of literature sets produced by different techniques or upon different networks. Instead of the standard core-periphery schemes, visualization may confront cognitive viewpoints, where areas of convergence (overlaps) are considered as strong forms (another form of core) and non-overlapping parts as possible extensions to be validated. An example of crossmaps was shown in Sect. 3.4. In the limit case of Boolean formulas addressing the whole domain to delineate, this would be equivalent to running a word-based search AND/OR a citation-based search. The AND clause yields the strong form and the OR clause a possible expansion along two branches, words and citations.

  The principle can be extended in a pragmatic way, given that (a) data analysis methods are not very robust and tend to yield quite different outcomes (b) data from different networks do not lead to identical results (poly-representation). Therefore the combination of methods, or the combination of networks, provides both ways to enhance precision ("strong

forms" where outcomes of different reliable methods converge), and ways to enhance recall, in divergence areas, at some risk. Zones of strong convergence can be considered "in."

- Decision Region and Cluster Evaluation (Fig. 3.8).
  - Evaluation at the cluster level. Again, thematic clusters are understood here in a broad meaning, whatever the data analysis method used. As a rule, there is no ground truth making the evaluation of recall, precision and F-scores or variants straightforward, so the relevance of each cluster has to be assessed by indirect indicators and/or supervision based on available cluster data. A manual light expertise uses cluster aggregate information such as label, pseudo-title recomposed from most specific words or phrases, ranked list of words, specific journals, cited authors/institutions, etc. Specificity of attributes is calculated by Tf-Idf or other indexes. Features from a previous IR or mapping process, say ratios of expansion to core, or results from crossmaps, are particularly helpful. Map displays using pleasant interfaces make the task easier.
  - Evaluation at finer granularity level. Finer-grain information can be available from the delineation protocol: IR projections of good quality onto a map; clusters crossings from hybrid methods; combination with zones of bibliometric expansion, etc. In such cases the border region may be treated at the infra-cluster or the document level. In pure mapping exercises, the cluster level may simply reveal too coarse, with exceedingly large or heterogeneous groupings. In this case, one has to go deeper in the cluster composition, through sampling for detailed analysis or further breakdown, at a cost.

The driving of evaluation is conditioned by the mastering of methodological effects and biases, likely to yield very different outputs. A particular attention, at the domain level, should be brought to the tendency of metrics and methods to favor particular semantic dimensions: to what extent can a domain be extended towards its intellectual base, especially theoretical foundations? Towards its tools and techniques? Towards its objects and products? Decision rules, in absence of a IR standard, will be based on quantitative indicators of the process, for example the intensity of bibliometric linkages in expansion stages, and experts' advices in terms of subjective precision, recall, and their balance (tantamount to variants of F-score). The convergence of experts' preferences, with the help of self-rating, may be taken into account.

3.5.2 To Conclude

Delineation at the meso-level deals with intermediary objects. Models in Price's tradition cast some light on the dynamics of the whole scientific system, whereas

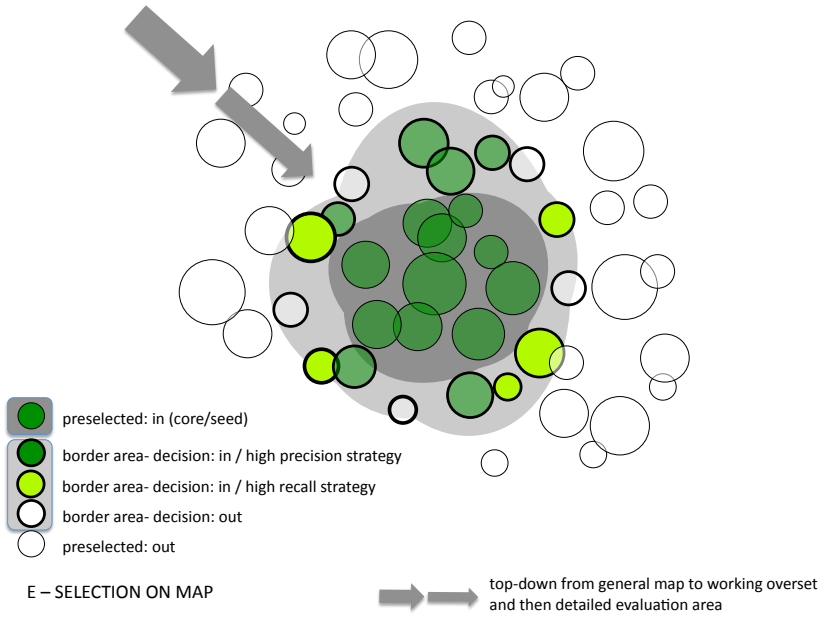


Figure 3.8: IR search and mapping approaches (3/3). (E) evaluation and decision on clusters in border area.

network theory proposes, at the micro-level, various mechanic models explaining emergence of meso-structures. The connection with practical solutions for topic and domain delineation, a rather multidisciplinary issue, will stimulate many research projects.

In practical studies, delineation operations should respect the proportionality principle. In simple cases, specialized and mature fields, the domain can be defined by using ready-made resources: official classifications, databases schemes. The complex cases which typically justify scientometric field studies — multidisciplinary, generic and emerging/unsettled domains — are precisely those where delineation and expertise are the more challenging. Coarse-grain approaches (journal-level) are easier to implement, but again hindered by locally complex network and abundance of non-specific media.

Bibliometrics both exploits and feeds science classification resources, literature searching and mapping models and human skill. Validation procedures include cross analyses and direct supervision. The delineation tasks pull together multiple strands of bibliometrics and IR. They inherit progress in data and network analysis, as well as common limitations in data coverage, robustness issues, ergonomy challenges with respect to supervision and discussions with sponsors. Bibliometrics cannot pretend to operationalize in a standard manner all questions from decision-makers nor, in cognitive applications, all questions from sociologists of science and other scholars.

Within the scope where "bibliometric hypothesis" applies, a horizon of de-

lineation is the comparison and combination of solutions from the networks which reflect scientific activity, essentially actors and institutions, citations and texts. Taking advantage of all available facets of data is a pragmatic choice, to which the concept of poly-representation has given a theoretical support. The cross-study of the three main universes associated to documents is also gaining attention in bibliometrics and sociology of research, supported by social network analysis. The theoretical profusion around models of growth and decline of communities is perhaps not settled now, but is very promising for understanding the invisible colleges in its various aspects. Will this multi-network research track converge towards unified hypotheses? There is little doubt that progress in this matter will enlighten the delineation issue especially in emerging areas. Meanwhile, the question remains whether networks should be fully hybridized with more or less radical techniques — substantive or featureless — or various networks solutions be conducted in parallel with final synthesis. In the background, the tremendous potential of deep learning on big science data is likely to reshuffle cards in retrieval and classification methods. The prospects are unclear right now, as their lack of explainability is a serious drawback in the bibliometric delineation context.

The management of supervision is central to the feasibility of bibliometric studies and their delineation tasks. Configurations are diverse, one cannot compare simple problems requiring light supervision, with large studies on controversial areas. In the latter case, the operators of the study deal with a possibly complex managerial organization, with steering committees and expert panels mixing policy makers, stakeholders and scientists, possibly with multiple roles. The selection of data sources and the methods of supervision, and finally the perimeter of the domain, will reflect those social stakes. The definition of fields or disciplines is particularly sensitive to academic interests, epistemic convictions and border issues, likely to create conflictual visions, sometimes between external observers and established players. The panel composition, to be efficient, should match the diversity of the domain, both in terms of thematic specialization and social stakes, with possibly some help from a few high-level generalists. In the mediation role, bibliometrics is also a social practice.

Bibliometric studies, if commissioned by administrations or institutions, enter a complex landscape of decision-help procedures where quantitative proposals are elements of discussion and decision among others. The question is vaster, however. Gläser et al. [26] underline the differences between operational definitions (say method outputs), pragmatic definitions (for clients and sponsors), and theoretical definitions (talking to science studies) of topics or domains. The notion of scientific domains is mobilized for a wide scope of purposes, labelling, information and evaluation in scientific institutions, science administration, IR databases of any kinds, laboratory life, scientists' self-positioning and last not least the reflexive work of scientometricians and social scientists on understanding the mechanisms of scientific activity.

Bibliography

[1] Auguste Comte: *Cours de philosophie positive*, Vol. 1 (Rouen Frères, Paris 1830)

[2] Robert K. Merton: Science and Technology in a Democratic Order, Journal of Legal and Political Sociology **1**(1), 115 – 126 (1942)

[3] Robert K. Merton: *The Sociology of Science: Theoretical and Empirical Investigations* (The University of Chicago Press, Chicago 1973)

[4] Thomas S. Kuhn: *The structure of scientific revolutions*, 2nd edn. (The University of Chicago Press, Chicago 1970)

[5] H. M. Collins, Steven Yearley: Epistemological Chicken. In: *Science as Practice and Culture*, ed. by Andrew Pickering (University of Chicago Press, Chicago 1992) Chap. 10, pp. 301 – 326

[6] Barry Barnes, David Bloor, John Henry: *Scientific knowledge: A sociological analysis* (The University of Chicago Press, Chicago 1996)

[7] David Bloor: *Knowledge and Social Imagery* (Routledge & Kegan Paul, London 1976)

[8] Karin D. Knorr-Cetina: Scientific Communities or Transepistemic Arenas of Research? A Critique of Quasi-Economic Models of Science, Social Studies of Science **12**(1), 101 – 130 (1982)

[9] Michael Joseph Mulkay, G. Nigel Gilbert, Steve Woolgar: Problem Areas and Research Networks in Science, Sociology **9**(2), 187 – 203 (1975)

[10] Michel Serres: *La Traduction*, Hermès III, Collection « Critique » (Les Éditions de Minuit, Paris 1974)

[11] Bruno Latour, Steve Woolgar: *Laboratory life: The social construction of scientific facts* (Sage Publications, Beverly Hills 1979)

[12] Michel Callon, Bruno Latour: Unscrewing the big Leviathan: how actors macro-structure reality and how sociologists help them to do so. In: *Advances in social theory and methodology: Toward an integration of mirco-*

and macro-sociologies, ed. by Karin Knorr-Cetina, Aaron Victor Cicourel (Routledge & Kegan Paul, Boston, London and Henley 1981) Chap. 10, pp. 277 – 303

[13] John Law, John Hassard: *Actor Network Theory and after* (Blackwell, Oxford 1999)

[14] Timothy Lenoir: *Instituting science: The cultural production of scientific disciplines* (Stanford University Press, Stanford 1997)

[15] Victor DiRita: Microbiology Is an Integrative Field, So Why Are We a Divided Society?, Microbe Magazine **8**(10), 384 – 385 (2013)

[16] Arturo Casadevall, Ferric C. Fang: Field Science—the Nature and Utility of Scientific Fields, mBio **6**(5), e01259 – 15 (2015)

[17] Jean Piaget: L'épistémologie des relations interdisciplinaires. In: *Interdisciplinarity: Problems of teaching and research in universities*, ed. by Léo Apostel, Guy Berger, Asa Briggs, Guy Michaud (OECD, Paris 1972) Chap. 1, pp. 127 – 140

[18] Derek John de Solla Price, Donald deB. Beaver: Collaboration in an invisible college, American Psychologist **21**(11), 1011 – 1018 (1966)

[19] Diana Crane: *Invisible colleges: Diffusion of knowledge in scientific communities* (Chicago University Press, Chicago 1972)

[20] Daryl E. Chubin: Beyond invisible colleges: Inspirations and aspirations of post-1972 social studies of science, Scientometrics **7**(3–6), 221 – 254 (1985)

[21] Alesia Zuccala: Modeling the invisible college, Journal of the American Society for Information Science and Technology **57**(2), 152 – 168 (2005)

[22] Jochen Gläser, Grit Laudel: Integrating Scientometric Indicators into Sociological Studies: Methodical and Methodological Problems, Scientometrics **52**(3), 411 – 434 (2001)

[23] Peter M. Haas: Introduction: Epistemic Communities and International Policy Coordination, International Organization **46**(1), 1 – 35 (1992)

[24] Étienne Wenger: *Communities of practice: Learning, meaning, and identity* (Cambridge University Press, New York 1998)

[25] Richard P. Smiraglia: Domain Analysis of Domain Analysis for Knowledge Organization: Observations on an Emergent Methodological Cluster, Knowledge Organization **42**(8), 602 – 611 (2015)

[26] Jochen Gläser, Andrea Scharnhorst, Wolfgang Glänzel: Same data—different results? Towards a comparative approach to the identification of thematic structures in science, Scientometrics **111**(2), 979 – 979 (2017)

[27] Cassidy R. Sugimoto, Scott Weingart: The kaleidoscope of disciplinarity, Journal of Documentation **71**(4), 775 – 794 (2015)

[28] Radosvet Todorov: Representing a scientific field: A bibliometric approach, Scientometrics **15**(5–6), 593 – 605 (1989)

[29] Robert J. W. Tijssen: A quantitative assessment of interdisciplinary structures in science and technology: Co-classification analysis of energy research, Research Policy **21**(1), 27 – 44 (1992)

[30] Caroline S. Wagner: *The new invisible college: Science for development* (Brookings Institution Press, Washington, DC 2008)

[31] Arho Suominen, Hannes Toivanen: Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification, Journal of the Association for Information Science and Technology **67**(10), 2464 – 2476 (2016)

[32] E. C. M. Noyons, A. F. J. van Raan: Monitoring scientific developments from a dynamic perspective: Self-organized structuring to map neural network research, Journal of the American Society for Information Science **49**(1), 68–81 (1998)

[33] Michel Zitt, Elise Bassecoulard: Delineating complex scientific fields by an hybrid lexical-citation method: An application to nanosciences, Information Processing & Management **42**(6), 1513 – 1531 (2006)

[34] Julie Thompson Klein: *Interdisciplinarity: History, theory, and practice* (Wayne State University Press, Detroit, MI 1990)

[35] Bernard C. K. Choi, Anita W. P. Pak: Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness., Clinical and Investigative Medicine **29**(6), 351 – 364 (2006)

[36] Thomas Jahn, Matthias Bergmann, Florian Keil: Transdisciplinarity: Between mainstreaming and marginalization, Ecological Economics **79**, 1 – 10 (2012)

[37] A. Wendy Russell, Fern Wickson, Anna L. Carew: Transdisciplinarity: Context, contradictions and capacity, Futures **40**(5), 460 – 472 (2008)

[38] Julie T. Klein: Evaluation of Interdisciplinary and Transdisciplinary Research, American Journal of Preventive Medicine **35**(2), S116 – S123 (2008)

[39] Thaddeus R. Miller, Timothy D. Baird, Caitlin M. Littlefield, Gary Kofinas, F. Stuart Chapin, III, Charles L. Redman: Epistemological pluralism: Reorganizing interdisciplinary research, Ecology and Society **13**(2), 46 (2008)

[40] Alfredo Yegros-Yegros, Ismael Rafols, Pablo D'Este: Does Interdisciplinary Research Lead to Higher Citation Impact? The Different Effect of Proximal and Distal Interdisciplinarity, PLOS ONE **10**(8), e0135095 (2015)

[41] Gregg E. A. Solomon, Stephen Carley, Alan L. Porter: How Multidisciplinary Are the Multidisciplinary Journals Science and Nature?, PLOS ONE **11**(4), e0152637 (2016)

[42] Cassidy R. Sugimoto, Nicolas Robinson-Garcia, Rodrigo Costas: Towards a global scientific brain: Indicators of researcher mobility using co-affiliation data, OECD Blue Sky III Forum on Science and Innovation Indicators, September 19-21, Ghent 2016, ed. by Maryann Feldman, Sadao Nagaoka, Luc Soete, Adam Jaffe, Monica Salazar, Reinhilde Veugelers (OECD, Paris 2016) online

[43] María Bordons, Fernanda Morillo, Isabel Gómez: Analysis of Cross-Disciplinary Research Through Bibliometric Tools. In: *Handbook of Quantitative Science and Technology Research: the Use of Publication and Patent Statistics in Studies of S&T Systems*, ed. by Henk F. Moed, Wolfgang Glänzel, Ulrich Schmoch (Kluwer, Dordrecht 2004) pp. 437 – 456

[44] Gabriel Pinski, Francis Narin: Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics, Information Processing & Management **12**(5), 297 – 312 (1976)

[45] Ed J. Rinia, Thed N. van Leeuwen, Eppo E. W. Bruins, Hendrik G. van Vuren, Anthony F. J. van Raan: Scientometrics **54**(3), 347 – 362 (2002)

[46] Elise Bassecoulard, Michel Zitt: Patents and Publications: The Lexical Connection. In: *Handbook of Quantitative Science and Technology Research: the Use of Publication and Patent Statistics in Studies of S&T Systems*, ed. by Henk F. Moed, Wolfgang Glänzel, Ulrich Schmoch (Kluwer, Dordrecht 2004) pp. 665 – 694

[47] Katy Börner, Richard Klavans, Michael Patek, Angela M. Zoss, Joseph R. Biberstine, Robert P. Light, Vincent Larivière, Kevin W. Boyack: Design and Update of a Classification System: The UCSD Map of Science, PLoS ONE **7**(7), e39464 (2012)

[48] Kevin W. Boyack, Richard Klavans: The Structure of Science. In: *Places & Spaces: Mapping Science — 1st Iteration (2005): The Power of Maps*, ed. by Katy Börner, Deborah MacPherson (scimaps.org, Indiana, IN 2005)

[49] Andy Stirling: A general framework for analysing diversity in science, technology and society, Journal of The Royal Society Interface **4**(15), 707 – 719 (2007)

[50] Diana Hicks: Limitations and More Limitations of Co-Citation Analysis/Bibliometric Modelling: A Reply to Franklin, Social Studies of Science **18**(2), 375 – 384 (1988)

[51] Henk F. Moed: *Citation Analysis in Research Evaluation*, Information Science and Knowledge Management, Vol. 9 (Springer, Dordrecht 2005)

[52] Anthony F. J. van Raan, Thed N. van Leeuwen, Martijn S. Visser: Severe language effect in university rankings: Particularly Germany and France are wronged in citation-based rankings, Scientometrics **88**(2), 495 – 498 (2011)

[53] Luc Soete, Susan Schneegans, Deniz Eröcal, Baskaran Angathevar, Rajah Rasiah: A world in search of an effective growth strategy. In: *UNESCO Science Report: Towards 2030*, UNESCO Reference Works, ed. by Susan Schneegans (UNESCO, Paris 2015) Chap. 1, pp. 20 – 55

[54] Jill Freyne, Lorcan Coyle, Barry Smyth, Padraig Cunningham: Relative status of journal and conference publications in Computer Science, Communications of the ACM **53**(11), 124 – 132 (2010)

[55] Anton J. Nederhof: Bibliometric monitoring of research performance in the Social Sciences and the Humanities: A Review, Scientometrics **66**(1), 81 – 100 (2006)

[56] Mu-hsuan Huang, Yu-wei Chang: Characteristics of research output in social sciences and humanities: From a research evaluation perspective, Journal of the American Society for Information Science and Technology **59**(11), 1819 – 1828 (2008)

[57] Gunnar Sivertsen, Birger Larsen: Comprehensive bibliographic coverage of the social sciences and humanities in a citation index: An empirical analysis of the potential, Scientometrics **91**(2), 567 – 575 (2012)

[58] Thed N. Van Leeuwen, Henk F. Moed, Robert J. W. Tijssen, Martijn S. Visser, Anthony F. J. Van Raan: Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance, Scientometrics **51**(1), 335 – 346 (2001)

[59] Michel Zitt, Suzy Ramanana-Rahary, Elise Bassecoulard: Correcting glasses help fair comparisons in international science landscape: Country indicators as a function of ISI database delineation, Scientometrics **56**(2), 259 – 282 (2003)

[60] Vincent Larivière, Éric Archambault, Yves Gingras, Étienne Vignola-Gagné: The place of serials in referencing practices: Comparing natural sciences and engineering with social sciences and humanities, Journal of the American Society for Information Science and Technology **57**(8), 997 – 1004 (2006)

[61] Carolin Michels, Ulrich Schmoch: The growth of science and database coverage, Scientometrics **93**(3), 831 – 846 (2012)

[62] Susanne Mikki: Comparing Google Scholar and ISI Web of Science for Earth Sciences, Scientometrics **82**(2), 321 – 331 (2010)

[63] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, Kuansan Wang: An Overview of Microsoft Academic Service (MAS) and Applications, WWW'15: Proceedings of the 24th International Conference on World Wide Web, Florence, Italy 2015, ed. by Aldo Gangemi, Stefano Leonardi, Alessandro Panconesi (ACM, New York 2015) 243 – 246

[64] Drahomira Herrmannova, Petr Knoth: An Analysis of the Microsoft Academic Graph, D-Lib Magazine **22**(9/10), online (2016)

[65] Anne-Wil Harzing, Satu Alakangas: Microsoft Academic: Is the phoenix getting wings?, Scientometrics **110**(1), 371 – 383 (2017)

[66] Jerry E. Gray, Michelle C. Hamilton, Alexandra Hauser, Margaret M. Janz, Justin P. Peters, Fiona Taggert: Scholar*ish*: Google Scholar and its Value to the Sciences, Issues in Science and Technology Librarianship **12**, online (2012)

[67] Cyril Labbé: Ike Antkare, one of the great stars in the scientific firmament, ISSI Newsletter **6**(2), 48 – 52 (2010)

[68] Péter Jacsó: Metadata mega mess in Google Scholar, Online Information Review **34**(1), 175 – 191 (2010)

[69] Anne-Wil Harzing, Satu Alakangas: Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison, Scientometrics **106**(2), 787 – 804 (2016)

[70] Qi Wang, Ludo Waltman: Large-scale analysis of the accuracy of the journal classification systems of Web of Science and Scopus, Journal of Informetrics **10**(2), 347 – 364 (2016)

[71] Mike Thelwall, Stefanie Haustein, Vincent Larivière, Cassidy R. Sugimoto: Do Altmetrics Work? Twitter and Ten Other Social Web Services, PLoS ONE **8**(5), e64841 (2013)

[72] Stefanie Haustein, Isabella Peters, Judit Bar-Ilan, Jason Priem, Hadas Shema, Jens Terliesner: Coverage and adoption of altmetrics sources in the bibliometric community, Scientometrics **101**(2), 1145 – 1163 (2014)

[73] Ehsan Mohammadi, Mike Thelwall: Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows, Journal of the Association for Information Science and Technology **65**(8), 1627 – 1638 (2014)

[74] Zohreh Zahedi, Rodrigo Costas, Paul Wouters: How well developed are altmetrics? A cross-disciplinary analysis of the presence of "alternative metrics" in scientific publications, Scientometrics **101**(2), 1491 – 1513 (2014)

[75] Carlos Luis González-Valiente, Josmel Pacheco-Mendoza, Ricardo Arencibia-Jorge: A review of altmetrics as an emerging discipline for research evaluation, Learned Publishing **29**(4), 229 – 238 (2016)

[76] Ann E. Williams: Altmetrics: An overview and evaluation, Online Information Review **41**(3), 311 – 317 (2017)

[77] Cinzia Daraio, Wolfgang Glänzel: Grand challenges in data integration—state of the art and future perspectives: An introduction, Scientometrics **108**(1), 391 – 400 (2016)

[78] OECD: *Revised Field of Science and Technology (FOS) Classification in the Frascati Manual — Report number DSTI/EAS/STP/NESTI(2006)19/FINAL* (OECD, Paris 2007)

[79] Eugene Garfield: The evolution of the Science Citation Index, International Microbiology **10**(1), 65 – 69 (2007)

[80] Alexander I. Pudovkin, Eugene Garfield: Algorithmic procedure for finding semantically related journals, Journal of the American Society for Information Science and Technology **53**(13), 1113 – 1119 (2002)

[81] Eugene Garfield: Citation Analysis as a Tool in Journal Evaluation: Journals can be ranked by frequency and impact of citations for science policy studies, Science **178**(4060), 471 – 479 (1972)

[82] Eugene Garfield: The History and Meaning of the Journal Impact Factor, Journal of the American Medical Association **295**(1), 90 – 93 (2006)

[83] Francis Narin, Gabriel Pinski, Helen Hofer Gee: Structure of the Biomedical Literature, Journal of the American Society for Information Science **27**(1), 25 – 45 (1976)

[84] Péter Jacsó: As we may search: Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases, Current Science **89**(9), 1537 – 1547 (2005)

[85] Félix de Moya-Anegón, Zaida Chinchilla-Rodríguez, Benjamín Vargas-Quesada, Elena Corera-Álvarez, Francisco José Muñoz-Fernández, Antonio González-Molina, Victor Herrero-Solana: Coverage analysis of Scopus: A journal metric approach, Scientometrics **73**(1), 53 – 78 (2007)

[86] Loet Leydesdorff, Susan E. Cozzens: The delineation of specialties in terms of journals using the dynamic journal set of the SCI, Scientometrics **26**(1), 135 – 156 (1993)

[87] Elise Bassecoulard, Michel Zitt: Indicators in a research institute: A multi-level classification of scientific journals, Scientometrics **44**(3), 323 – 345 (1999)

[88] Ismael Rafols, Martin Meyer: Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience, Scientometrics **82**(2), 263 – 287 (2009)

[89] Wolfgang Glänzel, András Schubert: A new classification scheme of science fields and subfields designed for scientometric evaluation purposes, Scientometrics **56**(3), 357 – 367 (2003)

[90] Eric Archambault, Olivier H. Beauchesne, Julie Caruso: Towards a multilingual, comprehensive and open scientific journal ontology, ISSI'11: Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics, Durban, South Africa 2011, ed. by Ed Noyons, Patrick Ngulube, Jacqueline Leta (ISSI, Leiden University and University of Zululand 2011) 66 – 77

[91] Kevin W. Boyack, Richard Klavans: Creation of a highly detailed, dynamic, global model and map of science, Journal of the Association for Information Science and Technology **65**(4), 670 – 685 (2014)

[92] Richard Klavans, Kevin W. Boyack: Which Type of Citation Analysis Generates the Most Accurate Taxonomy of Scientific and Technical Knowledge?, Journal of the Association for Information Science and Technology **68**(4), 984 – 998 (2017)

[93] Almudena Ruiz-Iniesta, Oscar Corcho: A review of ontologies for describing scholarly and scientific documents, SePublica'14: Proceedings of the the 4th Workshop on Semantic Publishing co-located with the 11th Extended Semantic Web Conference, Anissaras, Greece 2014, ed. by Alexander García Castro, Christoph Lange, Phillip Lord, Robert Stevens (CEUR-WS, Aachen 2014) online

[94] Alexander M. Petersen, Daniele Rotolo, Loet Leydesdorff: A triple helix model of medical innovation: Supply, demand, and technological capabilities in terms of Medical Subject Headings, Research Policy **45**(3), 666 – 681 (2016)

[95] Andrei Mogoutov, Bernard Kahane: Data search strategy for science and technology emergence: A scalable and evolutionary query for nanotechnology tracking, Research Policy **36**(6), 893 – 903 (2007)

[96] Alan L. Porter, Jan Youtie, Philip Shapira, David J. Schoeneck: Refining search terms for nanotechnology, Journal of Nanoparticle Research **10**(5), 715 – 728 (2007)

[97] Peter Ingwersen: Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory, Journal of Documentation **52**(1), 3 – 50 (1996)

[98] Jeppe Nicolaisen, Birger Hjørland: Practical potentials of Bradford's law: A critical examination of the received view, Journal of Documentation **63**(3), 359 – 377 (2007)

[99] Peter Ingwersen, Kalervo Järvelin: *The Turn: Integration of Information Seeking and Retrieval in Context*, The Information Retrieval Series, Vol. 18 (Springer, Dordrecht 2005)

[100] Thomas E. Nisonger: Journals in the Core Collection: Definition, Identification, and Applications, The Serials Librarian **51**(3–4), 51 – 73 (2007)

[101] Henry Small: Co-citation in the scientific literature: A new measure of the relationship between two documents, Journal of the American Society for Information Science **24**(4), 265 – 269 (1973)

[102] Quentin L. Burrell: On the h-index, the size of the Hirsch core and Jin's A-index, Journal of Informetrics **1**(2), 170 – 177 (2007)

[103] Wolfgang Glänzel, Bart Thijs: Using "core documents" for detecting and labelling new emerging topics, Scientometrics **91**(2), 399 – 416 (2012)

[104] J. Rocchio: Relevance Feedback in Information Retrieval. In: *The SMART retrieval system: Experiments in automatic document processing*, ed. by Gerard Salton (Prentice Hall, Englewood Cliffs, NJ 1971) pp. 313 – 323

[105] Gerard Salton, Chris Buckley: Improving retrieval performance by relevance feedback, Journal of the American Society for Information Science **41**(4), 288 – 297 (1990)

[106] Claudio Carpineto, Giovanni Romano: A Survey of Automatic Query Expansion in Information Retrieval, ACM Computing Surveys **44**(1), 1 – 50 (2012)

[107] Rakesh Agrawal, Tomasz Imieliński, Arun Swami: Mining association rules between sets of items in large databases, ACM SIGMOD Record **22**(2), 207 – 216 (1993)

[108] Darko Hric, Richard K. Darst, Santo Fortunato: Community detection in networks: Structural communities versus ground truth, Physical Review E **90**(6), 062805 (2014)

[109] Myer M. Kessler: Bibliographic coupling between scientific papers, American Documentation **14**(1), 10 – 25 (1963)

[110] Nick Jardine, Cornelis Joost van Rijsbergen: The use of hierarchic clustering in information retrieval, Information Storage and Retrieval **7**(5), 217 – 240 (1971)

[111] Philipp Mayr, Andrea Scharnhorst: Combining bibliometrics and information retrieval: Preface, Scientometrics **102**(3), 2191 – 2192 (2015)

[112] Philipp Mayr, Andrea Scharnhorst: Scientometrics and information retrieval: Weak-links revitalized, Scientometrics **102**(3), 2193 – 2199 (2015)

[113] Michel Zitt: Meso-level retrieval: IR-bibliometrics interplay and hybrid citation-words methods in scientific fields delineation, Scientometrics **102**(3), 2223 – 2245 (2015)

[114] Philipp Mayr, Ingo Frommholz, Guillaume Cabanac, Muthu Kumar Chandrasekaran, Kokil Jaidka, Min-Yen Kan, Dietmar Wolfram: Special Issue on Bibliometric-Enhanced Information Retrieval and Natural Language Processing for Digital Libraries, International Journal on Digital Libraries **19**, forthcoming (2018)

[115] M. E. J. Newman: The structure of scientific collaboration networks, Proceedings of the National Academy of Sciences **98**(2), 404 – 409 (2001)

[116] M. E. J. Newman: Coauthorship networks and patterns of scientific collaboration, Proceedings of the National Academy of Sciences **101**(Supplement 1), 5200 – 5205 (2004)

[117] Albert-László Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek: Evolution of the social network of scientific collaborations, Physica A: Statistical Mechanics and its Applications **311**(3–4), 590 – 614 (2002)

[118] Derek John de Solla Price: A general theory of bibliometric and other cumulative advantage processes, Journal of the American Society for Information Science **27**(5), 292 – 306 (1976)

[119] Réka Albert, Albert-László Barabási: Statistical mechanics of complex networks, Reviews of Modern Physics **74**(1), 47 – 97 (2002)

[120] Caroline S. Wagner, Loet Leydesdorff: Network structure, self-organization, and the growth of international collaboration in science, Research Policy **34**(10), 1608 – 1618 (2005)

[121] Gábor Csányi, Balázs Szendrői: Fractal–small-world dichotomy in real-world networks, Physical Review E **70**(1), 016122 (2004)

[122] Miller McPherson, Lynn Smith-Lovin, James M. Cook: Birds of a Feather: Homophily in Social Networks, Annual Review of Sociology **27**(1), 415 – 444 (2001)

[123] Nicolas Carayol, Pascale Roux: Knowledge flows and the geography of networks: A strategic model of small world formation, Journal of Economic Behavior & Organization **71**(2), 414 – 427 (2009)

[124] Katy Börner, Wolfgang Glänzel, Andrea Scharnhorst, Peter van den Besselaar: Modeling science: Studying the structure and dynamics of science, Scientometrics **89**(1), 347 – 348 (2011)

[125] Martine Cadot, Alain Lelu, Michel Zitt: Benchmarking 17 clustering methods. Available online at https://hal.archives-ouvertes.fr/hal-01532894 (2018)

[126] Andrew McCallum, Kamal Nigam, Lyle H. Ungar: Efficient clustering of high-dimensional data sets with application to reference matching, KDD'00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, MA 2000, ed. by Raghu Ramakrishnan, Sal Stolfo, Roberto Bayardo, Ismail Parsa (Association for Computing Machinery, New York, NY 2000) 169 – 178

[127] Michel Zitt, Elise Bassecoulard: Reassessment of co-citation methods for science indicators: Effect of methods improving recall rates, Scientometrics **37**(2), 223 – 244 (1996)

[128] Kevin W. Boyack, Richard Klavans: Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?, Journal of the American Society for Information Science and Technology **61**(12), 2389 – 2404 (2010)

[129] Glenn W. Milligan: A Review Of Monte Carlo Tests Of Cluster Analysis, Multivariate Behavioral Research **16**(3), 379 – 407 (1981)

[130] G. W. Milligan, M. C. Cooper: Methodology Review: Clustering Methods, Applied Psychological Measurement **11**(4), 329 – 354 (1987)

[131] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu: A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise, KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR 1996, ed. by Evangelos Simoudis, Jiawei Han, Usama Fayyad (AAAI, Palo Alto 1996) 226 – 231

[132] A. Rodriguez, A. Laio: Clustering by fast search and find of density peaks, Science **344**(6191), 1492 – 1496 (2014)

[133] Max Reinert: Un logiciel d'analyse lexicale, Les cahiers de l'analyse des données **11**(4), 471 – 481 (1986)

[134] Jean-Paul Benzécri: *L'analyse des correspondances*, Analyse des données, Vol. 2 (Dunod, Paris 1973)

[135] Peter D. Turney, Patrick Pantel: From frequency to meaning: vector space models of semantics, Journal of Artificial Intelligence Research **37**(1), 141 – 188 (2010)

[136] Scott Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, Beck L.: Improving information retrieval with latent semantic indexing, Proceedings of the 51st Annual Meeting of the American Society for Information Science **25**, 36 – 40 (1988)

[137] Alain Lelu: Clusters *and* factors: Neural algorithms for a novel representation of huge and highly multidimensional data sets. In: *New Approaches in Classification and Data Analysis*, ed. by Edwin Diday, Yves Lechevallier, Martin Schader, Patrice Bertrand (Springer, Berlin 1994) pp. 241 – 248

[138] Christos H. Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, Santosh Vempala: Latent semantic indexing: A probabilistic analysis, PODS'98: Proceedings of the 17th ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems, Seattle, WA 1998, ed. by Alberto Mendelson, Jan Paredaens (Association for Computing Machinery, New York, NY 1998) 159 – 168

[139] Thomas Hofmann: Probabilistic latent semantic indexing, SIGIR'99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, CA 1999, ed. by Fredric Gey, Marti Hearst, Richard Tong (ACM, New York, NY 1999) 50 – 57

[140] David M. Blei, Andrew Y. Ng, Michael I. Jordan: Latent Dirichlet Allocation, Journal of Machine Learning Research **3**, 993 – 1022 (2003)

[141] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre: Fast unfolding of communities in large networks, Journal of Statistical Mechanics: Theory and Experiment **2008**(10), P10008 (2008)

[142] Martin Rosvall, Carl T. Bergstrom: An information-theoretic framework for resolving community structure in complex networks, Proceedings of the National Academy of Sciences **104**(18), 7327 – 7331 (2007)

[143] Nees Jan van Eck, Ludo Waltman: Software survey: VOSviewer, a computer program for bibliometric mapping, Scientometrics **84**(2), 523 – 538 (2010)

[144] Marina Meila, Jianbo Shi: Learning Segmentation by Random Walks, NIPS'00: Proceedings of the Neural Information Processing Systems Conference, Denver, CO 2000, ed. by Todd K. Leen, Thomas G. Dietterich, Volker Tresp (MIT Press, Cambridge, MA 2000) 873 – 879

[145] Andrea Lancichinetti, Santo Fortunato: Community detection algorithms: A comparative analysis, Physical Review E **80**(5), 056117 (2009)

[146] Jure Leskovec, Kevin J. Lang, Michael Mahoney: Empirical comparison of algorithms for network community detection, WWW'10: Proceedings of the 19th international conference on World Wide Web, Raleigh, NC 2010,

ed. by Michael Rappa, Paul Jones, Juliana Freire, Soumen Chakrabarti (ACM, New York 2010) 631 – 640

[147] Jaewon Yang, Jure Leskovec: Defining and Evaluating Network Communities Based on Ground-Truth, ICDM'12: Proceedings of the 12th International Conference on Data Mining, Brussels 2012, ed. by Mohammed J. Zaki, Arno Siebes, Jeffrey Xu Yu, Bart Goethals, Geoff Webb, Xindong Wu (IEEE, Los Alamitos 2012) 745 – 754

[148] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, Gregoire Mesnil: A Latent Semantic Model with Convolutional-Pooling Structure for Information Retrieval, CIKM'14: Proceedings of the 23rd ACM conference on Information and knowledge mining, Shanghai 2014, ed. by Jianzhong Li, X. Sean Wang, Minos Garofalakis, Ian Soboroff, Torsten Suel, Min Wang (Association for Computing Machinery, New York, NY 2014) 101 – 110

[149] Christophe Van Gysel, Maarten de Rijke, Evangelos Kanoulas: Neural vector spaces for unsupervised information retrieval, arXiv Preprint **1708**(02702), online (2017)

[150] Tomas Mikolov, Wen tau Yih, Geoffrey Zweig: Linguistic Regularities in Continuous Space Word Representations, NAACL-HLT'13: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA 2013, ed. by Lucy Vanderwende, Hal Daume III, Katrin Kirchhoff (Association for Computational Linguistics, Stroudsburg, PA 2013) 746 – 751

[151] Omer Levy, Yoav Goldberg: Neural Word Embeddings as Implicit Matrix Factorization, NIPS'14: Proceedings of the Neural Information Processing Systems Conference, Montréal 2014, ed. by Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil Lawrence (Curran Associates, Red Hook, NY 2014) 2177 – 2185

[152] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, Mike Gatford: Okapi at TREC-3, TREC'94: Proceedings of the 3rd Text REtrieval Conference, Gaithersburg, MA 1994, ed. by Donna K. Harman (NIST, Gaithersburg, MA 1994) 109 – 126

[153] Thomas M. J. Fruchterman, Edward M. Reingold: Graph Drawing by Force-directed Placement, Software: Practice and Experience **21**(11), 1129 – 1164 (1991)

[154] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy: Gephi: An Open Source Software for Exploring and Manipulating Networks, ICWSM'09: Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media, San Jose, CA 2009, ed. by William W. Cohen, Nicolas Nicolov (AAAI, Palo Alto 2009) 361 – 362

[155] Shawn Martin, W. Michael Brown, Richard Klavans, Kevin W. Boyack: OpenOrd: An open-source toolbox for large graph layout, Proceedings of Visualization and Data Analysis 2011, San Francisco, CA 2011, ed. by Pak Chung Wong, Jinah Park, Ming C. Hao, Chaomei Chen, Katy Börner, David L. Kao, Jonathan C. Roberts (SPIE, Bellingham, WA 2011) 786806

[156] Wouter de Nooy, Aandrej Mrvar, Vladimir Batagelj: *Exploratory Social Network Analysis with Pajek*, Revised and Expanded 2nd edition edn. (Cambridge University Press, New York 2011)

[157] Martine Cadot, Alain Lelu: Optimized Representation for Classifying Qualitative Data, DBKDA'10: Proceedings of the 2nd International Conference on Advances in Databases, Knowledge, and Data Applications, Les Menuires, France 2010, ed. by Fritz Laux, Lena Strömbäck (IEEE, Los Alamitos 2010) 241 – 246

[158] Deng Cai, Xiaofei He, Jiawei Han: Document clustering using locality preserving indexing, IEEE Transactions on Knowledge and Data Engineering **17**(12), 1624 – 1637 (2005)

[159] William M. Rand: Objective Criteria for the Evaluation of Clustering Methods, Journal of the American Statistical Association **66**(336), 846 – 850 (1971)

[160] Thomas M. Cover, Joy A. Thomas: *Elements of Information Theory*, Wiley Series in Telecommunications, ed. by Donald L. Schilling (John Wiley & Sons, New York 1991)

[161] Peter Ronhovde, Zohar Nussinov: Multiresolution community detection for megascale networks by information-based replica correlations, Physical Review E **80**(1), 016109 (2009)

[162] Eugene Garfield, Alexander I. Pudovkin, V. S. Istomin: Why do we need algorithmic historiography?, Journal of the American Society for Information Science and Technology **54**(5), 400 – 412 (2003)

[163] Irena Marshakova: System of Document Connections Based on References, Nauchn-Techn. Inform. **6**(2), 3 – 8 (1973)

[164] Howard D. White, Belver C. Griffith: Author cocitation: A literature measure of intellectual structure, Journal of the American Society for Information Science **32**(3), 163 – 171 (1981)

[165] Gerard Salton: *The SMART retrieval system: Experiments in automatic document processing* (Prentice Hall, Englewood Cliffs, NJ 1971)

[166] Michel Callon, Jean-Pierre Courtial, William A. Turner, Serge Bauin: From translations to problematic networks: An introduction to co-word analysis, Social Science Information **22**(2), 191 – 235 (1983)

[167] William A. Turner, Ghislaine Chartron, F. Laville, B. Michelet: Packaging information for peer review: New co-word analysis techniques. In: *Handbook of Quantitative Science and Technology*, ed. by Anthony F. J. van Raan (Elsevier, Amsterdam 1988) Chap. 11, pp. 291 – 323

[168] John Whittaker: Creativity and Conformity in Science: Titles, Keywords and Co-word Analysis, Social Studies of Science **19**(3), 473 – 496 (1989)

[169] Linton C. Freeman: *The development of social network analysis:a study in the sociology of science* (Empirical Press, Vancouver, B. C. 2004)

[170] Chaomei Chen: CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature, Journal of the American Society for Information Science and Technology **57**(3), 359 – 377 (2006)

[171] Wolfgang Glänzel, Hans-Jürgen Czerwon: A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level, Scientometrics **37**(2), 195 – 221 (1996)

[172] Ludo Waltman, Nees Jan van Eck: A new methodology for constructing a publication-level classification system of science, Journal of the American Society for Information Science and Technology **63**(12), 2378 – 2392 (2012)

[173] Naoki Shibata, Yuya Kajikawa, Yoshiyuki Takeda, Katsumori Matsushima: Comparative study on methods of detecting research fronts using different types of citation, Journal of the American Society for Information Science and Technology **60**(3), 571 – 580 (2009)

[174] Bo Jarneving: A comparison of two bibliometric methods for mapping of the research front, Scientometrics **65**(2), 245 – 263 (2005)

[175] Katy Börner: *Atlas of Science: Visualizing What We Know* (MIT Press, Cambridge, MA 2010)

[176] Michel Zitt, Elise Bassecoulard: Development of a method for detection and trend analysis of research fronts built by lexical or cocitation analysis, Scientometrics **30**(1), 333 – 351 (1994)

[177] Loet Leydesdorff, Ismael Rafols: Interactive overlays: A new method for generating global journal maps from Web-of-Science data, Journal of Informetrics **6**(2), 318 – 332 (2012)

[178] Loet Leydesdorff, Ping Zhou: Nanotechnology as a field of science: Its delineation in terms of journals and patents, Scientometrics **70**(3), 693 – 713 (2007)

[179] Kevin W. Boyack: Investigating the effect of global data on topic detection, Scientometrics **111**(2), 999 – 1015 (2017)

[180] Carl Bergstrom: Eigenfactor: Measuring the value and prestige of scholarly journals, College & Research Libraries News **68**(5), 314 – 316 (2007)

[181] Michel Zitt, Henry Small: Modifying the journal impact factor by fractional citation weighting: The audience factor, Journal of the American Society for Information Science and Technology **59**(11), 1856–1860 (2008)

[182] Ludo Waltman, Nees Jan van Eck, Thed N. van Leeuwen, Martijn S. Visser: Some modifications to the SNIP journal impact indicator, Journal of Informetrics **7**(2), 272 – 285 (2013)

[183] Michel Zitt, Jean-Philippe Cointet: Citation impacts revisited: How novel impact measures reflect interdisciplinarity and structural change at the local and global level, ISSI'13: Proceedings of the 14th International Conference of the International Society for Scientometrics and Informetrics, Vienna, Austria 2013, ed. by Juan Gorraiz, Edgar Schiebel (Austrian Institute of Technology, Vienna 2013) 285 – 299

[184] Henry Small, E. Sweeney: Clustering the *Science Citation Index*® using co-citations: I. A comparison of methods, Scientometrics **7**(3–6), 391 – 409 (1985)

[185] Terttu Luukkonen, R. J. W. Tijssen, Olle Persson, Gunnar Sivertsen: The measurement of international scientific collaboration, Scientometrics **28**(1), 15 – 36 (1993)

[186] Michel Zitt, Elise Bassecoulard, Yoshiko Okubo: Shadows of the Past in International Cooperation: Collaboration Profiles of the Top Five Producers of Science, Scientometrics **47**(3), 627 – 657 (2000)

[187] Kevin W. Boyack, Richard Klavans, Katy Börner: Mapping the backbone of science, Scientometrics **64**(3), 351 – 374 (2005)

[188] Grant Lewison, Guillermo Paraje: The classification of biomedical journals by research level, Scientometrics **60**(2), 145 – 157 (2004)

[189] Simone Teufel, Jean Carletta, Marc Moens: An annotation scheme for discourse-level argumentation in research articles, EACL'99: Proceedings of the 9th Conference of the European chapter of the Association for Computational Linguistics, Bergen, Norway 1999, ed. by Henry S. Thompson, Alex Lascarides (ACL, Stroudsburg, PA 1999) 110 – 117

[190] Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, Dietrich Rebholz-Schuhmann: Automatic recognition of conceptualization zones in scientific articles and two life science applications, Bioinformatics **28**(7), 991 – 1000 (2012)

[191] Simone Teufel, Marc Moens: Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status, Computational Linguistics **28**(4), 409 – 445 (2002)

[192] Caroline Lyon, James Malcolm, Bob Dickerson: Detecting short passages of similar text in large document collections, EMNLP'01: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Pittsburgh, PA 2001, ed. by Lillian Lee, Donna Harman (ACL, Stroudsburg, PA 2001) 118 – 125

[193] Rudi Cilibrasi, Paul M. B. Vitányi: Clustering by Compression, IEEE Transactions on Information Theory **51**(4), 1523 – 1545 (2005)

[194] Charles H. Bennett, Péter Gács, Ming Li, Paul M. B. Vitányi, Wojciech H. Zurek: Information Distance, IEEE Transactions on Information Theory **44**(4), 1407 – 1423 (1998)

[195] Ming Li, Xin Chen, Xin Li, Bin Ma, Paul M. B. Vitányi: The Similarity Metric, IEEE Transactions on Information Theory **50**(12), 3250 – 3264 (2004)

[196] Rudi Cilibrasi, Paul M. B. Vitányi: The Google Similarity Distance, IEEE Transactions on Knowledge and Data Engineering **19**(3), 370 – 383 (2007)

[197] J. MacQueen: Some methods for classification and analysis of multivariate observations, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probabilities, Durban, South Africa 1967, ed. by Lucien M. Le Cam, Jerzy Neyman (University of California, Berkeley 1967) 281 – 297

[198] E. W. Forgy: Cluster analysis of multivariate data: Efficiency versus interpretability of classifications, Biometrics **21**, 768 – 769 (1965)

[199] Nachiketa Sahoo, Jamie Callan, Ramayya Krishnan, George Duncan, Rema Padman: Incremental hierarchical clustering of text documents, CIKM'06: Proceedings of the 15th ACM international conference on Information and knowledge management, Arlington, VA 2006, ed. by Philip S. Yu, Vassilis Tsotras, Edward Fox, Bing Liu (ACM, New York 2006) 357 – 366

[200] Hwanjo Yu, D. Searsmith, Xiaolei Li, Jiawei Han: Scalable Construction of Topic Directory with Nonparametric Closed Termset Mining, ICDM'04: Proceedings of the 4th IEEE International Conference on Data Mining, Brighton 2004, ed. by Rajeev Rastogi, Katharina Morik, Max Bramer, Xindong Wu (IEEE, Los Alamitos 2004) 1 – 4

[201] Fredrik Åström: Changes in the LIS research front: Time-sliced cocitation analyses of LIS journal articles, 1990–2004, Journal of the American Society for Information Science and Technology **58**(7), 947 – 957 (2007)

[202] David M. Blei, John D. Lafferty: Dynamic topic models, ICML'06: Proceedings of the 23rd international conference on Machine learning, Pittsburgh, PA 2006, ed. by William W. Cohen, Andrew Moore (ACM, New York 2006) 113 – 120

[203] Qiaozhu Mei, ChengXiang Zhai: Discovering evolutionary theme patterns from text: An exploration of temporal text mining, KDD'05: Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery and data mining, Chicago, IL 2005, ed. by Robert L. Grossman, Roberto Bayardo, Kristin Bennett, Jaideep Vaidya (Association for Computing Machinery, New York, NY 2005) 198 – 207

[204] Frizo Janssens, Wolfgang Glänzel, Bart De Moor: Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis, KDD'07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, San Jose, CA 2007, ed. by Pavel Berkhin, Rich Caruana, Xindong Wu, Scott Gaffney (Association for Computing Machinery, New York, NY 2007) 360 – 369

[205] Chaomei Chen, Fidelia Ibekwe-SanJuan, Jianhua Hou: The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis, Journal of the American Society for Information Science and Technology **61**(7), 1386 – 1409 (2010)

[206] David Chavalarias, Jean-Philippe Cointet: Phylomemetic Patterns in Science Evolution—The Rise and Fall of Scientific Fields, PLOS ONE **8**(2), e54847 (2013)

[207] Scott Shane: Technological Opportunities and New Firm Creation, Management Science **47**(2), 205 – 220 (2001)

[208] Kristina B. Dahlin, Dean M. Behrens: When is an invention really radical? Defining and measuring technological radicalness, Research Policy **34**(5), 717 – 737 (2005)

[209] Henry Small, Hung Tseng, Mike Patek: Discovering discoveries: Identifying biomedical discoveries using citation contexts, Journal of Informetrics **11**(1), 46 – 62 (2017)

[210] Eugene Garfield, Irving H. Sher: KeyWords Plus^{TM}—algorithmic derivative indexing, Journal of the American Society for Information Science **44**(5), 298 – 299 (1993)

[211] Ronald N. Kostoff, J. Antonio del Río, James A. Humenik, Esther Ofilia García, Ana María Ramírez: Citation mining: Integrating text mining and bibliometrics for research user profiling, Journal of the American Society for Information Science and Technology **52**(13), 1148 – 1156 (2001)

[212] Bart Verspagen, Claudia Werker: The Invisible College of The Economics of Innovation and Technological Change, Estudios de Economía Aplicada **21**(3), 187 – 203 (1975)

[213] Donald deB. Beaver, R. Rosen: Studies in scientific collaboration – Part III. Professionalization and the natural history of modern scientific co-authorship, Scientometrics **1**(3), 231 – 245 (1979)

[214] Terttu Luukkonen, Olle Persson, Gunnar Sivertsen: Understanding Patterns of International Scientific Collaboration, Science, Technology, & Human Values **17**(1), 101 – 126 (1992)

[215] Hildrun Kretschmer: Coauthorship networks of invisible colleges and institutionalized communities, Scientometrics **30**(1), 363 – 369 (1994)

[216] J. Sylvan Katz, Ben R. Martin: What is research collaboration?, Research Policy **26**(1), 1 – 18 (1997)

[217] J. S. Katz: Geographical proximity and scientific collaboration, Scientometrics **31**(1), 31 – 43 (1994)

[218] Jarno Hoekman, Koen Frenken, Robert J. W. Tijssen: Research collaboration at a distance: Changing spatial patterns of scientific collaboration within Europe, Research Policy **39**(5), 662 – 673 (2010)

[219] Theresa Velden, Asif-ul Haque, Carl Lagoze: A new approach to analyzing patterns of collaboration in co-authorship networks: Mesoscopic analysis and interpretation, Scientometrics **85**(1), 219 – 242 (2010)

[220] Peter Mutschke, Anabel Quan Haase: Collaboration and cognitive structures in social science research fields. Towards socio-cognitive analysis in information systems, Scientometrics **52**(3), 487 – 502 (2001)

[221] Julio Raffo, Stéphane Lhuillery: How to play the "Names Game": Patent retrieval comparing different heuristics, Research Policy **38**(10), 1617 – 1627 (2009)

[222] Katherine W. McCain: The author cocitation structure of macroeconomics, Scientometrics **5**(5), 277 – 289 (1983)

[223] G. Nigel Gilbert: Referencing as Persuasion, Social Studies of Science **7**(1), 113 – 122 (1977)

[224] Camille Roth, Jean-Philippe Cointet: Social and semantic coevolution in knowledge networks, Social Networks **32**(1), 16 – 29 (2010)

[225] Xavier Polanco, Luc Grivel, Jean Royauté: How to do things with terms in informetrics : terminological variation and stabilization as science watch indicators, ISSI'95: Proceedings of the 5th International Conference of the International Society for Scientometrics and Informetrics, River Forest, IL 1995, ed. by Michael E.D. Koenig, Abraham Bookstein (Learned Information, Medford NJ 1995) 435 – 444

[226] Martin F. Porter: An algorithm for suffix stripping, Program **14**(3), 130 – 137 (1980)

[227] Leo Egghe, Ronald Rousseau: *Introduction to informetrics: Quantitative methods in library, documentation, and information science* (Elsevier, Amsterdam 1990)

[228] Katherine W. McCain: Descriptor and citation retrieval in the Medical Behavioral Sciences literature: Retrieval overlaps and novelty distribution, Journal of the American Society for Information Science **40**(2), 110 – 114 (1989)

[229] Miranda Lee Pao: Term and citation retrieval: A field study, Information Processing & Management **29**(1), 95 – 112 (1993)

[230] Lutz Bornmann, Hans-Dieter Daniel: What do citation counts measure? A review of studies on citing behavior, Journal of Documentation **64**(1), 45 – 80 (2008)

[231] Blaise Cronin: *The Citation Process: The Role and Significance of Citations in Scientific Communication* (Taylor Graham, London 1984)

[232] Henry G. Small: Cited Documents as Concept Symbols, Social Studies of Science **8**(3), 327 – 340 (1978)

[233] Bruno Latour: *Science in Action: How to Follow Scientists and Engineers Through Society* (Harvard University Press, Cambridge, MA 1987)

[234] Alberto Cambrosio, Peter Keating, Simon Mercier, Grant Lewison, Andrei Mogoutov: Mapping the emergence and development of translational cancer research, European Journal of Cancer **42**(18), 3140 – 3148 (2006)

[235] Francis Narin, Eliott Noma: Is technology becoming science?, Scientometrics **7**(3-6), 369 – 381 (1985)

[236] Michel Callon: Pinpointing Industrial Invention: An Exploration of Quantitative Methods for the Analysis of Patents. In: *Mapping the Dynamics of Science and Technology*, ed. by Michel Callon, John Law, Arie Rip (Macmillan, Houndmills and London, UK 1986) Chap. 10, pp. 163 – 188

[237] Vincent Larivière, Éric Archambault, Yves Gingras: Long-term variations in the aging of scientific literature: From exponential growth to steady-state science (1900–2004), Journal of the American Society for Information Science and Technology **59**(2), 288 – 296 (2008)

[238] Ed C. M. Noyons, Renald K. Buter, Anthony F. J. van Raan, Holger Schwechheimer, Matthias Winterhager, Peter Weingart: *The Role of Europe in World-Wide Science and Technology: Monitoring and Evaluation in a Context of Global Competition — Report for the European Commission* (CWTS-Leiden and IWT-Bielefeld, Brussels 2000)

[239] Ed C. M. Noyons, R. K. Buter, Anthony F. J. van Raan, U. Schmoch, T. Heinze, S. Hinze, R. Rangnow: *Mapping Excellence in Science and Technology across Europe Nanoscience and Nanotechnology — Report of project EC-PPN CT-2002-0001 to the European Commission* (CWTS and Fraunhofer ISI, Leiden and Karlsruhe 2003)

[240] Michel Zitt, Alain Lelu, Elise Bassecoulard: Hybrid citation-word representations in science mapping: Portolan charts of research fields?, Journal of the American Society for Information Science and Technology **62**(1), 19 – 39 (2011)

[241] Patricia Laurens, Michel Zitt, Elise Bassecoulard: Delineation of the genomics field by hybrid citation-lexical methods: interaction with experts and validation process, Scientometrics **82**(3), 647–662 (2010)

[242] Steven A. Morris, Gary G. Yen: Crossmaps: Visualization of overlapping relationships in collections of journal papers, Proceedings of the National Academy of Sciences **101**(Supplement 1), 5291 – 5296 (2004)

[243] Cavan Reilly, Changchun Wang, Mark Rutherford: A rapid method for the comparison of cluster analyses, Statistica Sinica **15**(1), 19 – 33 (2005)

[244] Richard Klavans, Kevin W. Boyack: Toward a consensus map of science, Journal of the American Society for Information Science and Technology **60**(3), 455 – 476 (2009)

[245] Loet Leydesdorff, Ismael Rafols: A global map of science based on the ISI subject categories, Journal of the American Society for Information Science and Technology **60**(2), 348 – 362 (2009)

[246] Per Ahlgren, Bo Jarneving: Bibliographic coupling, common abstract stems and clustering: A comparison of two document-document similarity approaches in the context of science mapping, Scientometrics **76**(2), 273 – 290 (2008)

[247] Erjia Yan, Ying Ding: Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other, Journal of the American Society for Information Science and Technology **63**(7), 1313 – 1326 (2012)

[248] Theresa Velden, Kevin W. Boyack, Jochen Gläser, Rob Koopman, Andrea Scharnhorst, Shenghui Wang: Comparison of topic extraction approaches and their results, Scientometrics **111**(2), 1169 – 1221 (2017)

[249] Henry Small: Co-Citation Context Analayses And The Structure of Paradigms, Journal of Documentation **36**(3), 183 – 196 (1980)

[250] Henry Small: Maps of science as interdisciplinary discourse: Co-citation contexts and the role of analogy, Scientometrics **83**(3), 835 – 849 (2010)

[251] Simone Teufel, Advaith Siddharthan, Dan Tidhar: Automatic classification of citation function, EMNLP'06: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Sydney, Australia 2006, ed. by Dan Jurafsky, Éric Gaussier (ACL, Stroudsburg, PA 2006) 103 – 110

[252] Anna Ritchie, Stephen Robertson, Simone Teufel: Comparing citation contexts for information retrieval, CIKM'08: Proceeding of the 17th ACM conference on Information and knowledge mining, Napa Valley, CA 2008, ed. by James G. Shanahan, Sihem Amer-Yahia, Ioana Manolescu, Yi Zhang, David A. Evans, Alek Kolcz, Key-Sun Choi, Abdur Chowdury (Association for Computing Machinery, New York, NY 2008) 213 – 222

[253] Shengbo Liu, Chaomei Chen: The differences between latent topics in abstracts and citation contexts of citing papers, Journal of the American Society for Information Science and Technology **64**(3), 627 – 639 (2013)

[254] Henry Small: Interpreting maps of science using citation context sentiments: A preliminary investigation, Scientometrics **87**(2), 373 – 388 (2011)

[255] Aaron Elkiss, Siwei Shen, Anthony Fader, Güneş Erkan, David States, Dragomir Radev: Blind men and elephants: What do citation summaries tell us about a research article?, Journal of the American Society for Information Science and Technology **59**(1), 51 – 62 (2008)

[256] Alison Callahan, Stephen Hockema, Gunther Eysenbach: Contextual cocitation: Augmenting cocitation analysis and its applications, Journal of the American Society for Information Science and Technology **61**(6), 1130 – 1143 (2010)

[257] Xiaofeng He, C.H.Q. Ding, Hongyuan Zha, H.D. Simon: Automatic topic identification using webpage clustering, ICDM'01: Proceedings of the International Conference on Data Mining, San Jose, CA 2001, ed. by Nick Cercone, T. Y. Lin, Xindong Wu (IEEE, Los Alamitos 2001) 195 – 202

[258] Sergey Brin, Lawrence Page: The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems **30**(1–7), 107 – 117 (1998)

[259] Peter van den Besselaar, Gaston Heimeriks: Mapping research topics using word-reference co-occurrences: A method and an exploratory case study, Scientometrics **68**(3), 377 – 393 (2006)

[260] Per Ahlgren, Cristian Colliander: Document–document similarity approaches and science mapping: Experimental comparison of five approaches, Journal of Informetrics **3**(1), 49 – 63 (2009)

[261] Frizo Janssens, Wolfgang Glänzel, Bart De Moor: A hybrid mapping of information science, Scientometrics **75**(3), 607 – 631 (2008)

[262] Wolfgang Glänzel, Bart Thijs: Using "core documents" for the representation of clusters and topics, Scientometrics **88**(1), 297 – 309 (2011)

[263] Rob Koopman, Shenghui Wang, Andrea Scharnhorst: Contextualization of topics: Browsing through the universe of bibliographic information, Scientometrics **111**(2), 1119 – 1139 (2017)

[264] Yann LeCun: A Path to AI, BAI'17: Workshop on Beneficial Artificial Intelligence, Asilomar, CA 2017, ed. by Erik Brynjolfsson, Eric Horvitz, Peter Norvig, Francesca Rossi, Stuart Russell, Bart Selman, Max Tegmark (Future of Life Institute, Cambridge, MA 2017) https://futureoflife.org/wp-content/uploads/2017/01/Yann-LeCun.pdf

[265] Robert R. Braam, Henk F. Moed, Anthony F. J. van Raan: Mapping of science by combined co-citation and word analysis. I. Structural aspects, Journal of the American Society for Information Science **42**(4), 233–251 (1991)

[266] Birger Larsen: Exploiting citation overlaps for Information Retrieval: Generating a boomerang effect from the network of scientific papers, Scientometrics **54**(2), 155–178 (2002)

[267] Ying Huang, Jannik Schuehle, Alan L. Porter, Jan Youtie: A systematic method to create search strategies for emerging technologies based on the Web of Science: Illustrated for "Big Data", Scientometrics **105**(3), 2005–2022 (2015)