Scientometrics (2017) 111:1169–1221
DOI 10.1007/s11192-017-2306-1

# Comparison of topic extraction approaches and their results

**Theresa Velden[1,2] · Kevin W. Boyack[3] · Jochen Gläser[2] · Rob Koopman[4] · Andrea Scharnhorst[5] · Shenghui Wang[4]**

Received: 6 June 2016 / Published online: 7 March 2017


**Abstract** This is the last paper in the Synthesis section of this special issue on 'Same Data, Different Results'. We first provide a framework of how to describe and distinguish approaches to topic extraction from bibliographic data of scientific publications. We then compare solutions delivered by the different topic extraction approaches in this special issue, and explore where they agree and differ. This is achieved without reference to a ground truth, since we have to assume the existence of multiple, equally important, valid perspectives and want to avoid bias through the adoption of an ad-hoc yardstick. Instead, we apply different ways to quantitatively and visually compare solutions to explore their commonalities and differences and develop hypotheses about the origin of these differences. We conclude with a discussion of future work needed to develop methods for comparison and validation of topic extraction results, and express our concern about the

✉ Theresa Velden
velden@ztg.tu-berlin.de

Kevin W. Boyack
kboyack@mapofscience.com

Jochen Gläser
jochen.glaser@ztg.tu-berlin.de

Rob Koopman
rob.koopman@oclc.org

Andrea Scharnhorst
andrea.scharnhorst@dans.knaw.nl

Shenghui Wang
shenghui.wang@oclc.org

[1] University of Michigan School of Information, Ann Arbor, MI, USA

[2] Present Address: ZTG, Technical University Berlin, Hardenbergstr. 16-18, 10623 Berlin, Germany

[3] SciTech Strategies, Inc., 8421 Manuel Cia Pl NE, Albuquerque, NM 87122, USA

[4] OCLC Research, Schipholweg 99, Leiden, The Netherlands

[5] DANS-KNAW, Anna van Saksenlaan 51, The Hague, The Netherlands

lack of access to non-proprietary benchmark data sets to support method development in the field of scientometrics.

**Keywords** Topic extraction · Comparative methods · Astrophysics · Data modeling · Clustering · Topic labeling · Science mapping

## Introduction

Topic extraction from scientific literature seems to be as much an art as a science. Different teams within the field of scientometrics use different approaches, based on their familiarity with specific methods, investment in the development of specific tools, long-term experience with the mapping of scientific fields, and in-house experimentation to optimize an approach. Rarely results are published that apply alternative approaches to the same data set and compare the results, and there is a lack of understanding of how differences between approaches affect the results obtained. In what ways do the solutions that they produce differ from one another? Is one approach better than another? What are the knobs and levers' of each approach, and how do they affect the results? As laid out in the introduction of this special issue (Gläser et al. 2017) there is a growing need to have some certainty about to what extent structures emerging from methodological approaches are indeed representation of thematic structures in science or artifacts produced by methods themselves.

To shed light on these questions, we have applied to the same data set a variety of topic extraction approaches that are documented in articles in this special issue. The data set consists of bibliographic data of documents in the astrophysics literature and is hereafter called the *Astro Data Set*. In this article we provide a comparative overview of the properties of these approaches and the topic solutions that they deliver. However, due to the fluidity of cognitive structures in science, and the multiplicity of reference frames (Gläser et al. 2017), there is no single ground truth that would tell us authoritatively how to divide the documents in the *Astro Data Set* (or any other set of scholarly articles) into topics. Therefore, how to compare the topic solutions and generate useful descriptions of their differences in the presence of multiple, inaccessible ground truths, is a research problem in its own right. For our purposes here, we will be interested in descriptions of solutions and their differences that:

- Capture various dimensions of how a solution differs from other solutions;
- Reveal the distinctiveness of the perspective that a solution provides into the topical structure of the field;
- Generate hints for differences between solutions that can be attributed to specific properties of the approaches used.

Ideally, our comparisons would be reviewed by area experts, who could evaluate the merits of the different perspectives created by the solutions [a further research direction discussed in Gläser et al. (2017)]. Meanwhile, we were fortunate to have some knowledge expertise within our author team with several authors trained in engineering or physics, including a subarea of astrophysics. We also occasionally discussed our results with astrophysicists outside of the author group. This allowed us to bring to bear this expertise on the interpretation of topical structures that were constructed by the various approaches. But from the point of view of reproducibility of the results and their interpretation this can

also be seen as problematic. We used two ways to at least acknowledge this problem: (a) by being transparent and articulating whenever subject expertise guided the analysis and (b) by trying to find reproducible ways to compare the interpretative dimensions (e.g. in the labeling of structures) across approaches.

In this paper we provide a first insight into how topic extraction approaches construct topics differently and how the resulting topical structures differ. Eventually, we would like to deliver guidance to the scientometrics community and users of topic extraction results on how to choose among approaches and what to keep in mind when interpreting results. More work needs to be done that we describe in the final section of this paper.

Comparing approaches, detecting in what aspects results agree or differ, and trying to understand why, is the core of this paper. The paper proceeds as follows: First, we provide a framework to characterize the approaches, and discuss where differences in their work flows arise ("Overview on topic extraction approaches" section). Second, we introduce methods that we used to compare topic solutions ("Tools for comparing topic extraction solutions" section). Third, an ensemble-based comparison of the solutions they generate is conducted ("Findings: comparisons across whole solutions" section). This comparison by means of an overview is countered by a number of specific comparisons. They are guided by assumptions about which perspective of the self-organised nature of the emergence of scientific topics is placed in the foreground by each of the methods ("Findings: specific comparisons" section). The paper concludes with a discussion of how the discourse around comparison of methods and approaches could be further fostered in the scientometric community.

## The data set

The *Astro Data Set* consists of the bibliographic data of 111,616 publications published in the years 2003–2010 in 59 astrophysical journals indexed by Web of Science (see "Data set: journal titles" in Appendix for the list of journal titles). To cover primarily original scientific content only documents of type Article, Letter, and Proceedings Paper were included, whereas document types Biographical-Item, Book Review, Correction, Editorial Material, Meeting Abstract, News Item, Reprint and Review were excluded. Reference links between publications were reconstructed by matching bibliographic information using a rule-based script developed by Michael Heinz (Humboldt University).

## Overview on topic extraction approaches

The selection of the eight topic extraction approaches compared in this paper is opportunistic in that these are the approaches developed and used by the teams that have come together to collaborate and produce this special issue on 'Same Data, Different Results'.[1] This means that for each approach used in this comparison there is one member or team in our collaboration who is intimately familiar with that approach. What occurred in the discussions at a series of workshops over a couple of years, is to which extent each of us had to make informed, and sometimes pragmatic, decisions on what approach to pursue and how to tweak it to meet the specific objectives of our respective research and tool development projects. These discussions led to a framework or language to characterize

[1] See the introduction of this special issue by Gläser et al. (2017) for the history and purpose of this collaboration.

**Table 1** Combinations of data models and clustering algorithms

| | Direct citation | Bibliographic coupling (bc) | Hybrid (bc + NLP terms) | Semantic matrix | Global direct citation map |
|---|---|---|---|---|---|
| Infomap | u | – | – | – | – |
| SLMA | c | – | – | – | sr |
| Memetic | hd | – | – | – | – |
| Louvain | – | eb | en | ol | – |
| k-means | – | – | – | ok | – |

and distinguish features of approaches, including the distinction between a 'data modeling' component and a 'clustering algorithm' component that is reflected in the organization of Table 2 which provides an overview of the distinguishing properties of the approaches and of the specific solutions that we decided to include in our comparison.[2]

Table 1 provides an overview of the various combinations of data models and clustering algorithms covered by the solutions included in our comparison (and what areas of the potential space of combinations are left unexplored due to resource limitations). The three solutions *c* (Van Eck and Waltman 2017), *hd* (Havemann et al. 2017), and *u* (Velden et al. 2017) are delivered by a set of approaches that model the data as a direct citation network, but use different clustering algorithms; another two solutions, *eb* and *en* are delivered by a set of approaches that use the same clustering algorithm, but model the data slightly differently: the first one as a bibliographic coupling network and the second one as a hybrid network based on bibliographic coupling in combination with terms extracted using Natural Language Processing (NLP) (Glänzel and Thijs 2017); another set of approaches models the data as a semantic matrix by interpreting each bibliographic or other metadata field as a semantic entity and applies two different clustering algorithms (Koopman and Wang 2017a), delivering solutions *ol* and *ok*, respectively; finally, solution *sr* (Boyack 2017a) is generated by using the direct citation network of a superset of literature from a global science map and projecting the *Astro Data Set* onto a clustered version of that map. All eight topic extraction approaches and their results are described in detail in the corresponding companion articles in this special issue.

The way the data is modeled (what features of the articles in the data set are extracted and used to represent the data) and the choice of clustering algorithm that is used to detect regularities in the data and extract groups of articles that represent candidates for 'topics' are key differences between approaches. Importantly for our purpose here, the set of approaches in this special issue covers a wide range of ways to model data and a number of clustering algorithms. But, there are clearly also dimensions missed, as for instance, author relations. Note further that all approaches in this sample use a document (or link) clustering algorithm. Future work should include also topic modeling approaches and possibly hybrid document clustering and topic modeling approaches, such as Xie and Xing (2013). Still, the variety within our sample makes it suitable as a first set to explore the question of how

[2] Some of the teams produced several solutions at different levels of resolution. Due to resource constraints we decided to limit the number of solutions included in the comparison. When selecting the solutions to include in our comparison we sought to limit deviations due to extreme differences in resolution. e.g. from the four solutions offered by Van Eck and Waltman (2017) with 22, 42, 115, 434 clusters respectively, we chose the solution with 22 clusters to coincide with the 22 clusters delivered by Velden et al. (2017). The two sets of solutions offered by Glänzel and Thijs (2017) were all of very low resolution, hence we chose the solutions with highest resolution: a solution with 13 instead of the alternative solution with 6 clusters, and a solution with 11 instead of the alternative with 5 clusters.

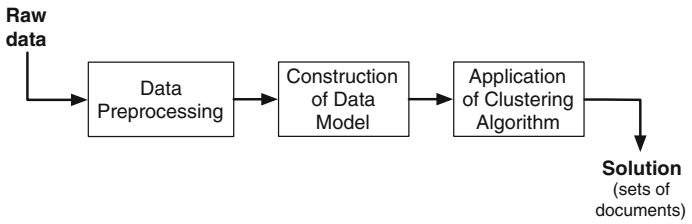



**Fig. 1** Schematic of a typical topic extraction workflow. Topic extraction approaches can differ in any of these steps, thereby producing variation between solutions

approaches and their results differ. The data models cover citation based models, hybrid models (citation and text based), and so-called semantic models. The algorithms used include four of the most popular clustering algorithms, namely k-means (MacKay 2003), Infomap (Rosvall and Bergstrom 2008), Louvain (Blondel et al. 2008), and Smart Local Moving Algorithm (Waltman and Eck 2012, 2013), which is an improved variant of the Louvain algorithm, along with a new memetic type algorithm (Havemann et al. 2017). The latter has been designed specifically for the extraction of overlapping, poly-hierarchical topics in the scientific literature.

Figure 1 schematically depicts the steps of a typical topic extraction work flow that consists of data preprocessing, construction of a data model, and the selection and application of a clustering algorithm. Differences between approaches can occur along any of these steps. Whereas for the purpose of this comparison all teams start from the same data set of source documents ("The data set" section), a first source of divergence during the preprocessing of this raw data is that some teams proceed by mapping this data set to their in-house database (Boyack 2017a; Glänzel and Thijs 2017). As some publications cannot be mapped to an entity in those in-house databases, those teams work with smaller subsets of documents. Also the information contained in those in-house databases on each publication (e.g. information about references to other publications) may differ from the information used by those teams that worked with the original data set. To give an example, the team that provided solutions *en* and *eb* had access to the unique reference codes given by Thompson Reuters to construct citation links between documents (Glänzel and Thijs 2017), whereas other participants worked with the reference links deducted from rule base parsing of reference strings mentioned in "The data set" section.

A fundamental difference between the solutions produced by the various approaches is their coverage, ranging from 91–100% of the 111, 616 documents in the *Astro Data Set*. The reason for this variation can be found not only in differences in the preprocessing stage of the data, but also during further steps in the workflow, as described in the following:

*Solutions ok and ol* As can be seen in Table 2, the most comprehensive solutions are *ok* and *ol* with a coverage of 100%, delivering 31 topics and 32 topics respectively.

*Solutions en and eb* Next in terms of coverage are solutions *en* and *eb* that include 97.99 and 97.22% of all documents, delivering 11 topics and 13 topics, respectively. These solutions were generated from a data set that was created by mapping the *Astro Data Set* onto an in-house version of the Web of Science, which resulted in reduction of the original set to 110,412 publications ($\sim$99%) (Glänzel and Thijs 2017). This subset was further reduced in two steps. First, in the data modeling step 82 documents were excluded from *en* because they did not reach a chosen similarity threshold for the minimal lexical similarity between any two documents in the data set. For *eb* the data modeling step resulted in 1479 documents being dropped because they did not share any references with any other

**Table 2** Properties of approaches

| Label | Data model | Clustering | Parameters | Coverage (%) | Topics |
|---|---|---|---|---|---|
| c | Direct citation Giant component | Smart Local Moving Algorithm (SLMA) | Resolution Min. cluster size | 101, 828 (91.23%) | 22 |
| u | [*Same as above*] | Infomap (undirected) | Random seed | 101, 831 (91.23%) | 22 |
| ol | Semantic matrix (network of 40 most similar neighbors) | Louvain (python library networkX) | Word occurrence thresh.<br>Stopword list<br>K most similar articles<br>Similarity value thresh | 111, 616 (100%) | 32 |
| ok | Semantic matrix | k-means (python library sklearn.cluster. MiniBatchKMeans) | Word occurrence thresh.<br>Stopword list<br>Number of clusters | 111, 616 (100%) | 31 |
| eb | Bibliographic coupling (bc) | Louvain (pajek) | References <11 years if in TR product<br>Resolution<br>Link strength thresh. | 108,512 (97.22%) | 13 |
| en | Bc and lexical coupling (NLP) | Louvain (pajek) | Resolution<br>Link strength thresh.<br>Weight bc versus text | 109,376 (97.99%) | 11 |
| sr | Direct citation incl. non-source items cited $\geq 2$ | Projection onto 1996–2012 global science map (SLMA) | Resolution<br>Min. cluster size | 107,304 (96.14%) | 555 |
| hd | Direct citation Giant component | Memetic (random evolution + deterministic search) | Seeds<br>Resolution<br>Population size<br>Other evolution param. | 101,762 (91.17%) | 111 over-lapping |

documents in the data set (and hence did not couple). Second, after the clustering step, all documents were excluded from solutions *en* and *eb* that had been assigned by the clustering algorithm to single document clusters or 'small, irrelevant' clusters (Glänzel and Thijs 2017), resulting in 954 documents getting omitted from solution *en* with a final coverage of 110,330 documents, and 421 documents getting omitted from solution *eb* with a final coverage of 108,512 documents.

*Solution sr* Solution *sr* delivers 555 topics. During data preprocessing, the source data was mapped to an in-house database from SCOPUS, resulting in a reduction to 107,888 documents (96.66% of the full *Astro Data Set*). The clustering step consisted of locating those remaining documents in the global map of science clustered at the region-level (Boyack 2017a). In this step, 584 documents could not be located in the global science map indicating that they were not included in the creation of the global map because of missing reference or citation information. Therefore the final *sr* solutions covers 107,304 documents in total which corresponds to a coverage of 96.14%.

*Solutions u, c and hd* Solutions *c*, *u*, and *hd* have the lowest coverage. They are based on the direct citation model and include only the documents in the giant component of the direct citation network.[3] Solutions *c* and *u* both deliver 22 topics and their coverage is

[3] All other connected components were omitted, since the largest of those was only 48 documents in size and considered too small to constitute a topic.

nearly the same at 91.23%. Solution *c* omits three documents that are connected to the giant component of the direct citation network only by future pointing references, whereas those three documents are included in *u*. Solution *hd* has a slightly smaller coverage (91.17%) and delivers 111 topics. It covers 66 documents less than solution *u* due to an additional selection process after the clustering step: Out of a total set of 381 valid clusters produced by the approach only a subset of 113 clusters was selected to meet criteria for a minimum cluster size of 20 papers and a minimum quality of clusters as measured by the associated cost function (see Havemann et al. 2017, for details). We further decided to include only 111 of the 113 clusters and omit the two largest clusters of this solution from the comparison as they provided only limited information about the topical structure of the *Astro Data Set*.[4]

Finally, a number of parameters usually need to be set in the modeling of the data and in the application of a clustering algorithm that influence the results achieved, such as a minimum threshold for the strengths of links to be considered in a bibliographic coupling network, or a requirement for a minimum size of clusters to be extracted. In Table 2 we list those parameters for each approach.

## Tools for comparing topic extraction solutions

We use a variety of tools to compare solutions and capture differences in how they group documents from the *Astro Data Set* into topics. We use a quantitative measure to get a first idea about the similarity or disparity of solutions. We use visual mappings of solutions onto various reference frames that support comparing solutions to one another. Finally, we explored a variety of labeling approaches to capture the content of a topic and compare solutions with regard to the content of the topics that they construct.

### Metrics: Normalized Mutual Information

To quantify the degree of similarity between solutions, we used an information theoretic measure that is commonly used in computer science to compare clusterings, namely Normalized Mutual Information (NMI). It considers membership in a cluster as a random variable and quantifies to what extent knowing one clustering reduces uncertainty about the other clustering. See "Comparison metric: Normalized Mutual Information" in Appendix for details on how this measure is defined.

### Labeling

*Thesaurus terms* Our first approach to labeling clusters makes use of thesaurus terms from the Unified Astronomy Thesaurus (UAT), a public domain thesaurus specific to astronomy.[5] As described in detail in Boyack (2017b), it contains 1915 unique terms at a

[4] These two clusters, respectively, represented 63 and 73% of the entire data. Their overlap was large with 50,684 documents and their union essentially included the entire subset of the 101,831 documents in the data model of the *hd* approach. Hence these two clusters contribute only limited information about the topical structure of the *Astro Data Set*. The ten largest clusters of the remaining 111 clusters have sizes 71,488, 62,606, 31,242, 19,948, 17,372, 11,899, 10,406, 9035, 8201, 4990, and their union covers 90% of the total data set.

[5] http://astrothesaurus.org.

maximum depth of 12 levels. The *Astro Data Set* was indexed to generate thesaurus terms for each document using title and abstract as input.[6] To generate cluster labels we used the most specific terms assigned to each document plus level 2 terms. Of these we selected as labels the most relevant terms as determined by a NMI measure that compares distribution of terms in one cluster with that in other clusters (see Koopman and Wang 2017 for details). See "Data files" in Appendix for download information for the corresponding data file.

*Natural language terms* A second approach to labeling clusters used terms extracted from the titles and abstracts of documents. As for the thesaurus terms, we constructed labels by selecting the ten terms with highest NMI scores when cluster documents are compared to non-cluster documents. This labeling approach is described in detail by Koopman and Wang (2017b). The labels for all clusters $\geq$100 documents are given in "Cluster-level labels" of Appendix. See also "Data files" of Appendix for download information for the corresponding data file.

*Journal signature* In Velden et al. (2017) a high-level classification of document clusters is introduced that builds on the observation that groups of clusters share similarities with regard to their most popular and distinctive journal titles ('journal signature'). Using this approach, six scientific domains were distinguished that seem to correspond to sub-disciplines within Astronomy and Astrophysics: Gravitation and Cosmology, Astroparticle Physics, Astrophysics, Solar Physics, Planetary Science, Space Science. Based on their journal signature, the 35 largest clusters of each of the seven disjoint cluster solutions that are included in our comparison, were assigned to those domains. Most assignments were straightforward, but some cases were ambiguous and more difficult to decide when a journal signature exposed a mixture of characteristics. This high-level grouping of clusters by scientific domains provides yet another reference frame for comparisons, as solutions differ in how they divide up a domain into topics, and how they shape the interfaces between domains. See "Data files" in Appendix for download information for the corresponding data files.

## Visual mapping

*Little Ariadne* As described in Koopman et al. (2017), *Little Ariadne* is a special instantiation of *Ariadne*, a user friendly tool for browsing bibliographic databases. This specific instance uses the bibliographic information in the *Astro Data Set* and is available at http://thoth.pica.nl/astro/. In our analysis we use the tool to visualize how the document clusters provided by the eight different approaches relate to one another in an abstract semantic space. Similarity here is based on a semantic matrix that is created from indexing entities such as authors, journals (ISSN), subjects, citations, topical terms, MAI thesaurus terms, cluster IDs, and citations (see Koopman et al. 2017 for details). The visualization we produce with *Little Ariadne* highlights which clusters from different solutions are very similar to one another, and which solutions produced clusters that are relatively distinct from all clusters produced by the other solutions.

*Lexical fingerprint* The lexical fingerprint is a method to quantify and visually compare the topical content of individual clusters, within a solution and across all solutions (Koopman and Wang 2017b). It builds on the mutual information based labeling of document clusters described above. The lexical terms that constitute the baseline of the

[6] Indexing was done by Access Innovations using their MAI (Machine Aided Indexer) software package and the UAT rule base that they maintain.

fingerprint are selected in a two step process: First, for each solution a ranked list of the 50 terms with highest NMI score is created. Then a joint set of terms for the fingerprint is created, by selecting the 50 highest ranking terms across those lists, excluding terms that appear only on one solution's list. For the visualization of the fingerprint of a cluster, the joint list of 50 terms is arranged along the *x*-axis based on their similarity according to the semantic matrix used for *Little Ariadne*. The *y*-axis gives the NMI score of a cluster for the respective terms. The resulting lexical fingerprints looks like radiation emission spectra, except that the values on the *x*-axis do not represent continuous values of wavelengths or frequencies but instead are terms, and hence categorical values. See "Data files" in Appendix for download information for the corresponding data file with a list of fingerprint terms and the scores for all clusters $\geq$100 documents from all eight solutions.

*Affinity networks* The construction of topic affinity networks is a method to map and visualize the internal structure of a solution. The method shows how the document clusters extracted relate to one another based on direct citation links between documents. In the calculation of link strengths between document clusters only the surplus of citations relative to a random null model (based on cluster sizes) are considered in order to reduce the 'cluttering' of the visualization from a pervasive background of connectivity within the scientific literature (see Velden et al. 2017; Velden and Lagoze 2013, for details of the method). See "Data files" in Appendix for download information for the corresponding data files.

## Findings: comparisons across whole solutions

### Differences in topic size distributions

Figure 2 shows the accumulative size distribution of the document clusters that are extracted by the eight approaches. Given the overlap of clusters in the *hd* solution, we removed duplicates from unions of clusters when calculating the accumulative fractional size. The distribution shows that solutions *hd*, *sr* and *en* are highly concentrated in that they reach a coverage of 75% of the *Astro Data Set* by their first six largest clusters alone.[7] By contrast, solutions *ol* and *ok* show much lower concentration, reaching 75% coverage only when including the 18 (*ol*) and 20 (*ok*) largest clusters, respectively.

### Degree of similarity between solutions

To get a first idea of the degree of similarity between solutions we use Normalized Mutual Information as a quantitative measure of the similarity between a pair of solutions (see Table 3). Note that this metric as well as the topic affinity networks used further below, could only be produced for disjoint cluster solutions such that *hd* is excluded from the comparison in this section.

[7] We found a strong correlation of cluster concentration with the proportion of unique document pairs of a solution, that is those pairs of documents that are clustered together by only one solution. This is a consequence of large clusters allowing for many more combinations of papers into pairs than smaller clusters, such that solutions with a high concentration of papers in a few large clusters have disproportionally many unique pairs. We had hoped that proportion of unique pairs could be an insightful measure of the distinctiveness of perspective that a particular solutions provides relative to all other solutions in the comparison, but we had to recognize that the distinctiveness in perspective that is signaled by proportion of unique pairs is primarily its cluster concentration.

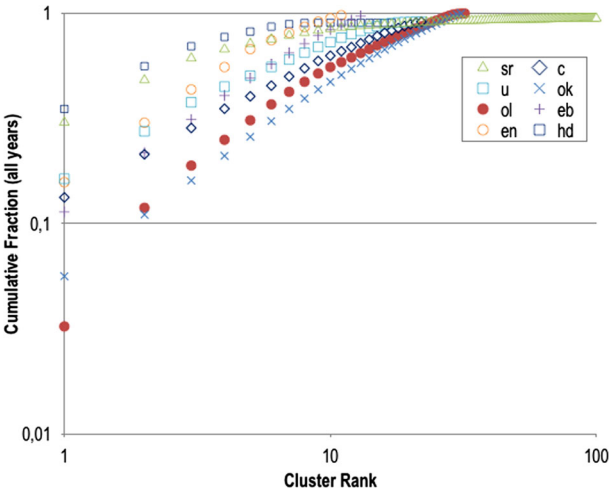



**Fig. 2** Accumulative fractional size distribution of clusters in each solution. The *y*-axis indicates what fraction of the total set of 111,616 documents are included, the *x*-axis corresponds to cluster rank, ordered by cluster size

**Table 3** Normalised Mutual Information (emphasis: **max**, *min* value)

| | sr | c | u | ok | ol | en | eb |
|---|---|---|---|---|---|---|---|
| sr | 1.00 | 0.36 | 0.37 | 0.33 | 0.33 | *0.24* | 0.31 |
| c | 0.36 | 1.00 | 0.63 | 0.46 | 0.52 | 0.32 | 0.38 |
| u | 0.37 | **0.63** | 1.00 | 0.42 | 0.47 | 0.30 | 0.36 |
| ok | 0.33 | 0.46 | 0.42 | 1.00 | 0.52 | 0.33 | 0.36 |
| ol | 0.33 | 0.52 | 0.47 | 0.52 | 1.00 | 0.31 | 0.36 |
| en | 0.24 | 0.32 | 0.30 | 0.33 | 0.31 | 1.00 | 0.33 |
| eb | 0.31 | 0.38 | 0.36 | 0.36 | 0.36 | 0.33 | 1.00 |

The median of the distribution of NMI scores is 0.36. The highest similarity score of NMI=0.63 is obtained for the pair of solutions *c* and *u*. Both are based on the same data model, have nearly identical coverage of data, and differ only in the clustering algorithm used. Figure 3 shows groupings of solutions at different levels of agreement with respect to the quartile of the NMI score between each pair of solutions (1st quartile: NMI $< 0.32$, 2nd quartile: $0.32 \leq$ NMI $< 0.36$, 3rd quartile: $0.36 \leq$ NMI $< 0.44$, 4th quartile: NMI $\geq 0.44$). For example, the similarity score of each possible pairing of solutions in the set (*c*, *ok*, *eb*, *en*) is larger than the 1st quartile of similarity scores, i.e. NMI $\geq 0.32$. Besides the pair of solutions with the maximal NMI score, we find two overlapping groups of solutions with high similarity scores above the third quartile level (NMI $\geq 0.44$), namely (*u*, *c*, *ol*) and (*c*, *ol*, *ok*). In the following we will refer to the union of these two groups as the 'core group' of solutions. The next similar solutions to this core group are *sr* and *eb*. Solution *en* is the most dissimilar. It joins a subset of the core set only if we allow for NMI values as low as the 2nd quartile.

To visually inspect the degree of similarity between the solutions we generate their topic affinity networks (see Figs. 4, 5). An affinity network shows how the different topics

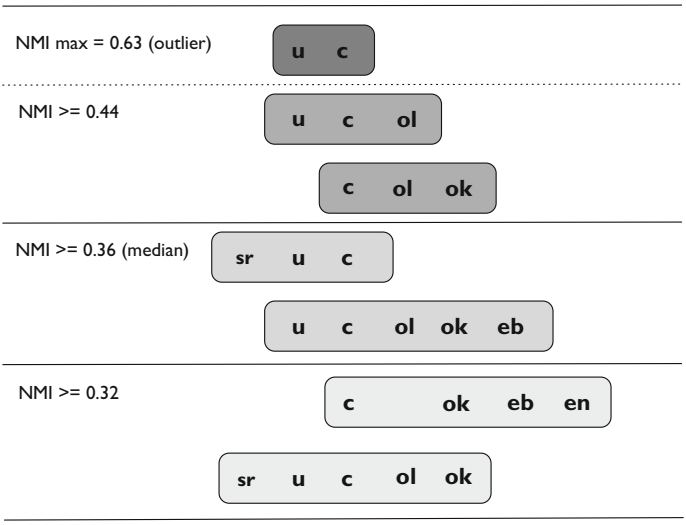



**Fig. 3** Grouping of clustering solutions based on degree of mutual similarities in cluster membership measured by NMI

within a solution connect to one another based on direct citations and thereby allows to visualize the topical structure that a solution imposes on the *Astro Data Set*. To support the comparison and interpretation of these maps,[8] we subdivide the affinity network into scientific domains based on the journal signature of the document clusters that constitute the nodes of the network (Velden et al. 2017).

The affinity networks in Figs. 4 and 5 reveal that in all seven solutions, the domain *Astrophysics* is the largest domain and most central domain in the sense that it interfaces with each of the other domains. Its relative size ranges between 50 and 55% of the documents covered by a solution, with the exception of *en* where it includes only about 42% of documents. Interestingly, the neighboring domain of *Planetary Science* is much larger in *en* than in all other solutions, suggesting that a number of documents that other solutions have assigned to clusters in the domain of *Astrophysics* may have been assigned by *en* to the *Planetary Science* domain instead (see our detailed investigation further below).

The topology of affinity networks in Figs. 4 and 5 underscores the similarity between the core group of solutions (*u*, *c*, *ol* and *ok*) that was already indicated by the quantitative NMI measure. It consists of an elongated structure with the domains of *Gravitational Physics and Cosmology* and *Astroparticle Physics* located at one end, the domain of *Astrophysics* in the middle, and the domains of *Solar Physics*, *Planetary Science*, and *Space Science* located at the other end. This structure suggests an organization of the field from objects at large scales of space-time and larger distance from earth to smaller objects and closer distance to earth, as discussed in Velden et al. (2017). Solutions *eb* and *sr* can be

[8] A note of caution: as usual with network visualizations, each map as represented is a projection of a multi-dimensional space onto a 2-dimensional space using one of many equally plausible layouts for the network. The network algorithm used optimizes readability of the map by avoiding overlap of nodes and crossing of edges. Different representations are possible and equally valid. Whether the node of a topic is on the left or right of the network, at the bottom or top of the network is arbitrary and not significant. The most relevant feature for interpreting the map is whether nodes are linked or not and the strength of the links.

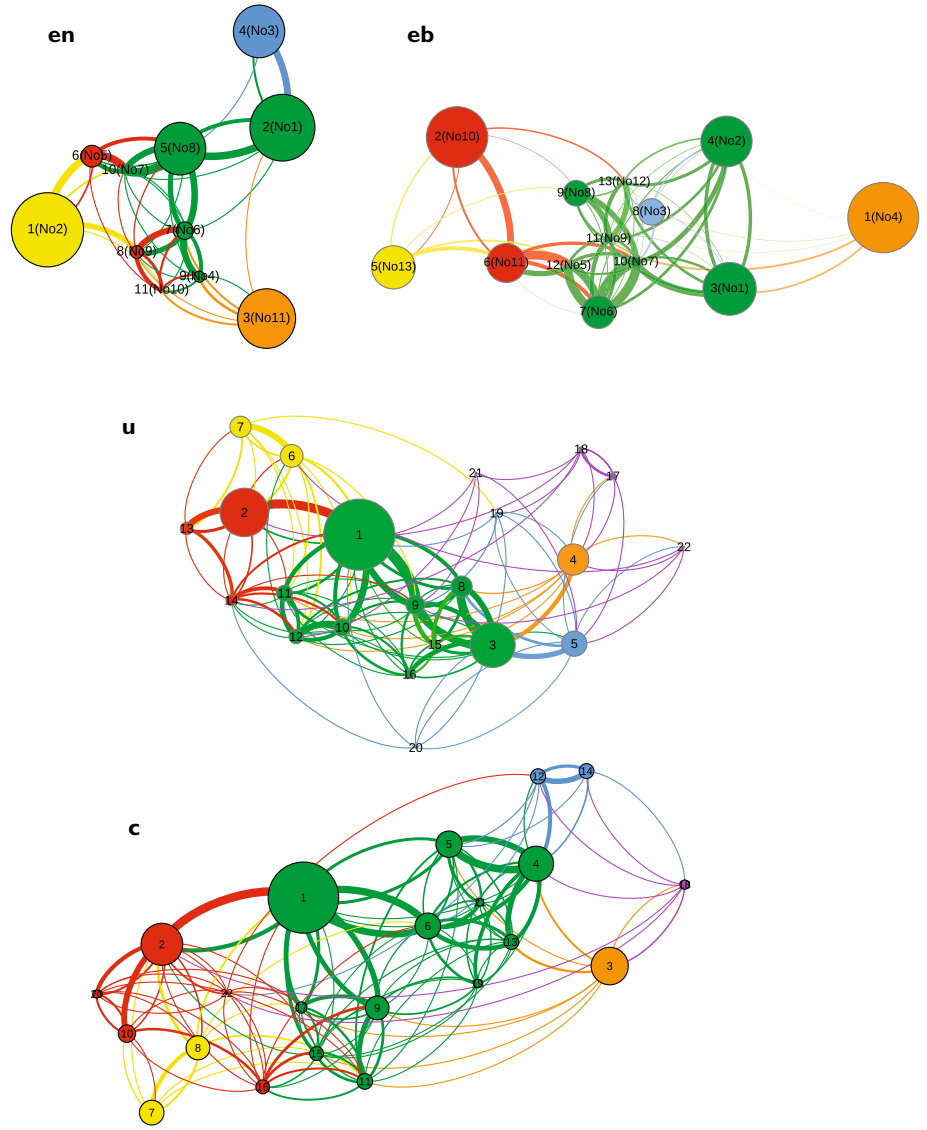



**Fig. 4** Topic affinity networks for solutions *en, eb, u, c*. *Node size* indicates number of documents, and link strength relative preference given by publications in one topic to cite publications in the other. Links are directed, *colored* by their source node and curve clockwise away from it. *Node colors* visible in the online version—indicate a scientific domain based on journal signature: *red* (Gravitational Physics and Cosmology), *yellow* (Astroparticle Physics), *green* (Astrophysics), *orange* (Solar Physics), *blue* (Planetary Science), *purple* (Space Science). The *first number in node labels* indicates rank of node by size, and the *second number, in brackets* are the cluster indeces provided by the creators of solutions and used in the remainder of the article to identify clusters [Network visualization: gephi + *Force Atlas* 2 algorithm, one of the few network layout algorithms that considers edge weights in directed networks]. (Color figure online)

seen as exposing variants of this pattern. Due to their very different number of clusters (13 versus 51[9]) they sit at opposite ends of the spectrum of solutions.

Solution *eb* shows a structure that is similar to the core group with the *Astrophysics* domain at the center. However, the low number of clusters in *eb* seemingly suppresses the separate identification of the domain of *Space Science*. An inspection of cluster labels provided in "Cluster-level labels" of Appendix reveals that topics that in other solutions are a core component of the domain of *Space Science*, such as those relating to 'solar wind' and 'ionosphere' (see cluster labels for e.g. c18, u17, u18, ok4, ok25) are included in *eb* in the *Solar Physics* domain.

Solution *sr* is distinct because of its much higher number of clusters (51), an extreme variation in cluster sizes, and its high concentration of documents in a small number of clusters (see also Fig. 2). Interestingly, the cluster size distribution in *sr* differs significantly across domains: whereas *Astrophysics*, *Solar Physics* and *Gravitational Physics and Cosmology* are each dominated by one or two large topics, the domains of *Planetary Science*, *Space Science* and *Astroparticle Physics* show significant scatter with a large number of small topics.

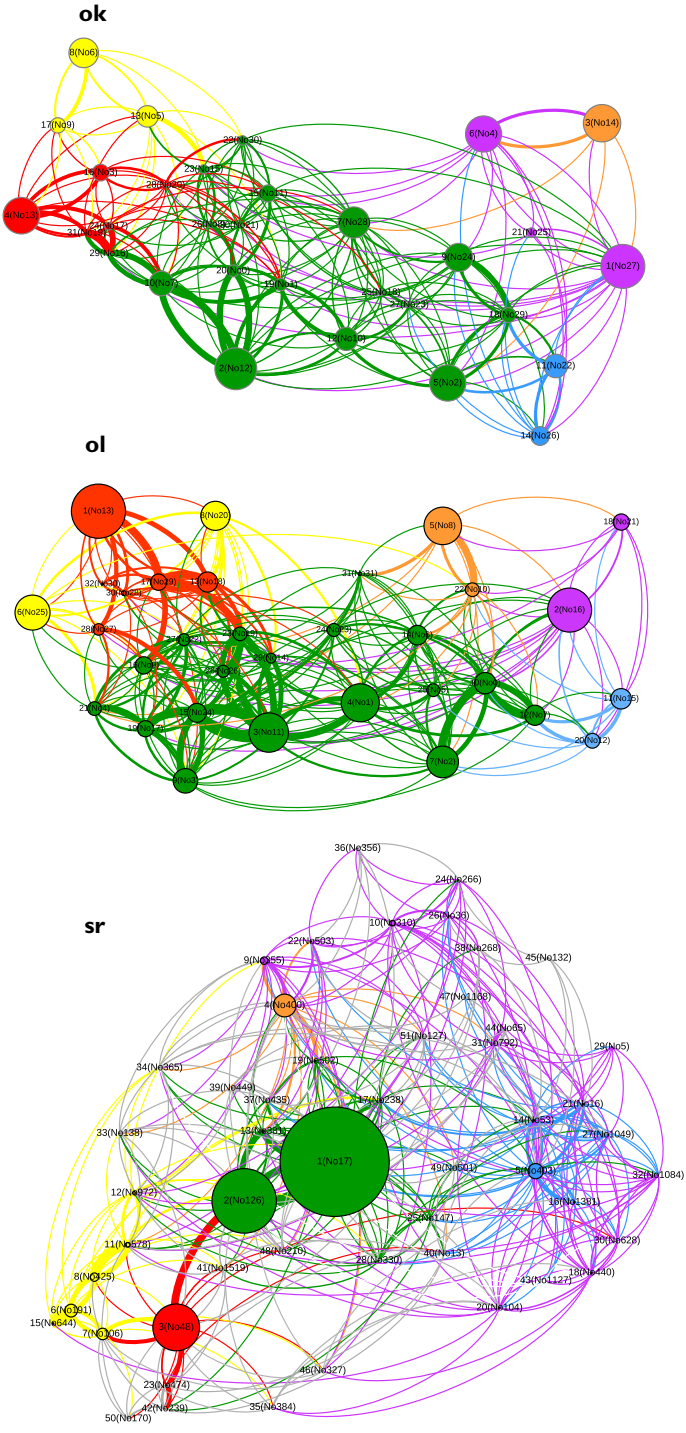
The topology of connections between domains in the affinity network of *sr* is similar to that of the core group of solutions: *Astrophysics* takes a central position and the other domains are split into two groups, with one group attaching to one end (*Gravitational Physics, Cosmology* and *Astroparticle Physics*), and the other group of domains attaching to the other end (*Solar Physics*, *Planetary Science*, and *Space Science*). The quantitative measure of similarity, NMI, suggests a relatively high (3rd quartile level) similarity between solutions *u*, *c*, and *sr* but not between *sr* and *ol* or *ok*. Looking at Figs. 4 and 5 this seems plausible because the former three solutions share a high concentration of documents in large clusters in the domains of *Astrophysics* and *Gravitational Physics and Cosmology* and a lack of such a concentration in the domain of *Space Science*. These tendencies are not shared by the *ol* and *ok* solutions that show greater scatter of documents across several clusters in *Astrophysics* and *Gravitational Physics and Cosmology*, and a concentration of documents in one or two larger clusters for the domain of *Space Science*.

Finally, the affinity network of solution *en* looks very distinct from the affinity networks of the other solutions. Besides having only a small number of clusters (11) and no topics assigned to the *Space Science* domain (two features it shares with the *eb* solution), *en* constructs the interface between *Solar Physics* and the other domains in a unique way. It links *Solar Physics* with *Gravitational Physics and Cosmology* through topic en10 ('gravitational waves') and with *Astrophysics* through topics en4 ('x-ray') and en6 ('gamma ray'), that both also interface directly with *Gravitational Physics and Cosmology*. This moves *Solar Physics* away from the other end of the elongated structure that characterizes the core group of solutions. This striking topological difference in combination with the low NMI similarity score of *en* when compared to any of the other solutions suggests a difference in the aggregation of topics that will be further investigated in "Citation based versus semantic data models" section.

## What solutions agree about

The visualization tool *Little Ariadne* can be used to generate a global view on all eight solutions and how their topic clusters relate to one another based on semantic similarity

[9] For detailed analysis and visualization purposes we restrict solution *sr* to the 51 clusters that are at least 50 documents in size.


ok
8(No6)
17(No9)
19(No5)
6(No4)
3(No14)
4(No13)
1(No27)
2(No12)
5(No21)
14(No26)
ol
1(No13)
5(No8)
6(No25)
2(No16)
4(No1)
3(No11)
7(No2)
sr
36(No356)
24(No266)
22(No503)
9(No355)
45(No132)
47(No1108)
4(No400)
29(No5)
34(No365)
39(No449)
1(No17)
2(No126)
12(No972)
11(No578)
3(No48)
15(No544)
50(No170)
35(No384)

◀**Fig. 5** Topic affinity networks for solutions *ok, ol, sr*. *Node size* indicates number of documents, and link strength relative preference given by publications in one topic to cite publications in the other. Links are directed, colored by their source node and curve clockwise away from it. *Node colors* visible in the online version—indicate a scientific domain based on journal signature: *red* (Gravitational Physics and Cosmology), *yellow* (Astroparticle Physics), *green* (Astrophysics), *orange* (Solar Physics), *blue* (Planetary Science), *purple* (Space Science). The *first number in node labels* indicates rank of node by size, and the *second number, in brackets*, are the cluster indeces provided by the creators of solutions and used in the remainder of the article to identify clusters [Network visualization: gephi + *Force Atlas* 2 ]. (Color figure online)

(see Fig. 6). Relationships between topics are based on the distance measure of the semantic matrix used by *Little Ariadne*. Eye-catching is the large-scale structure of this map with areas of higher concentration of topics and relative voids in between. This large-scale structure corresponds well to the high-level domains that were derived from journal signatures (see "Labeling" section). This suggests that the similarity in journal signatures

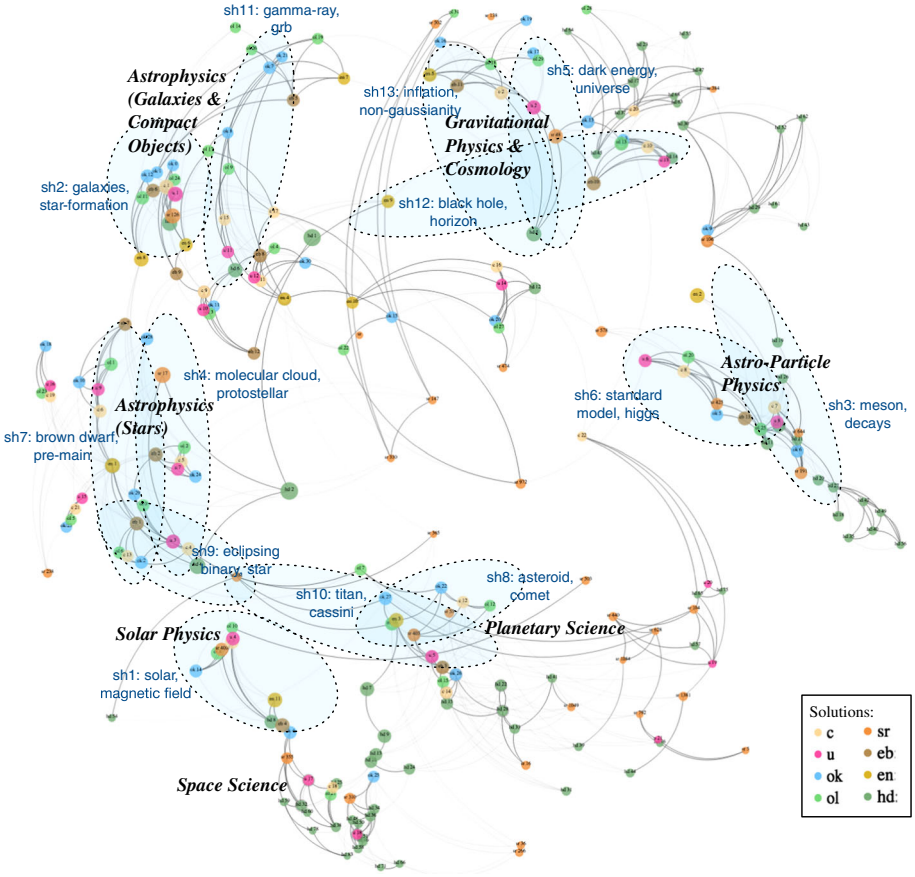



**Fig. 6** Relationships between clusters from all eight solutions as seen by Little Ariadne. The *bold* labels indicate high-level scientific domains (Velden et al. 2017) that correspond well to the large scale structure of the network of clusters shown here. The *dotted-line ovals* indicate the approximate location of the clusters that are associated with each of the 13 largest shared document sets (sh1–sh13). The shared document sets are labeled by the top two terms generated using the entropy based labeling method introduced in Koopman and Wang (2017a)

correlates to a large extent with semantic similarity as measured by Ariadne, with one caveat, namely that Fig. 6 suggests a subdivision of the *Astrophysics* domain into larger objects (galaxies) versus smaller objects (stars), which is a distinction that is not obvious in the analysis of the journal signatures of the corresponding topics.

To further explore the agreement between solutions, we looked for sets of documents that are clustered together into a single topic by every solution. Of the 111,616 documents in the data set, 96,921 are included in all solutions. There are 4289 (maximal) document sets, that include at least 2 documents and for which each solution has at least one cluster containing each set. We call these 'shared document sets' and interpret them as representing 'hard thematic cores' of documents that all solutions agree belong together in one topic. The 13 largest shared document comprise 23,217 documents and account for about 21% of the documents in the *Astro Data Set*. Their size and associated clusters are listed in "Shared document sets" of Appendix. The approximate position of the associated clusters in the cluster network is indicated in Fig. 6 by light blue ovals. With exception of the domain of *Space Science* that is not represented in solutions *eb* and *en*, all domains contain one or several of the 13 shared document sets, meaning they have thematic cores that are identified unambiguously by all eight topic extraction approaches. Labels for the shared document sets that describe the content of these thematic cores are given in Table 6 in "Shared document sets" of Appendix.

Upon inspection of the lists of clusters associated with each of the 13 largest shared document sets, we noticed two instances where the majority of solutions place two document sets into the same cluster whereas a small set of solutions disagrees and separates those two document sets into distinct topics. The first case concerns document sets 5 and 13 in the *Gravitational Physics and Cosmology* domain. A visual analysis of the lexical fingerprints of the clusters (see Fig. 7) shows that solutions *ok* and *ol* distinguish between the topics of 'inflation' and 'dark energy', whereas all other solutions combine these two topics into one. From a theoretical perspective, inflation (early universe expansion) and dark energy (current phase expansion) are separate phenomena, however they are potentially linked, which is the concern of the so called 'quintessence' theory in astrophysics. This suggests that from a subject expert's standpoint detecting the linkage between the topics as well as detecting the distinctiveness of the two topics provide informative perspectives on the topical structure of the field.

The second case concerns the domain of *Planetary Science* where solutions *c, ok, ol* assign the shared document sets 8 and 10 to two distinct topics. The set of clusters in solutions *c, ok, ol* that include document set 10 (ol15, ok26, c14) show a clear signal for the terms 'mars' and 'surface' (see Fig. 8). The set of clusters in solutions *c, ok, ol* relating to document set 8 (ol12, ok22, c12) however has only a very weakly expressed fingerprint. This is likely because the most relevant terms for these clusters were suppressed in the construction of the lexical space for the fingerprint analysis, an issue discussed in Koopman and Wang (2017a)). Consulting the cluster labels given for these three clusters in "Cluster-level labels" of Appendix, we find 'asteroid', and 'comet' listed as top terms for those clusters, terms that are not included in the lexical fingerprint. The labels for the shared document set 8 (see Table 6 in "Shared document sets" of Appendix) confirm that the topic of this second shared document set is focused on 'comets' and 'asteroids'. From an astrophysical perspective a distinction between research on asteroids and comets on the one hand (document set 8) and research on planets in the solar system (document set 10) seems a plausible one to make and whether to merge the two topics into a more general planetary science one would seem a matter of resolution.

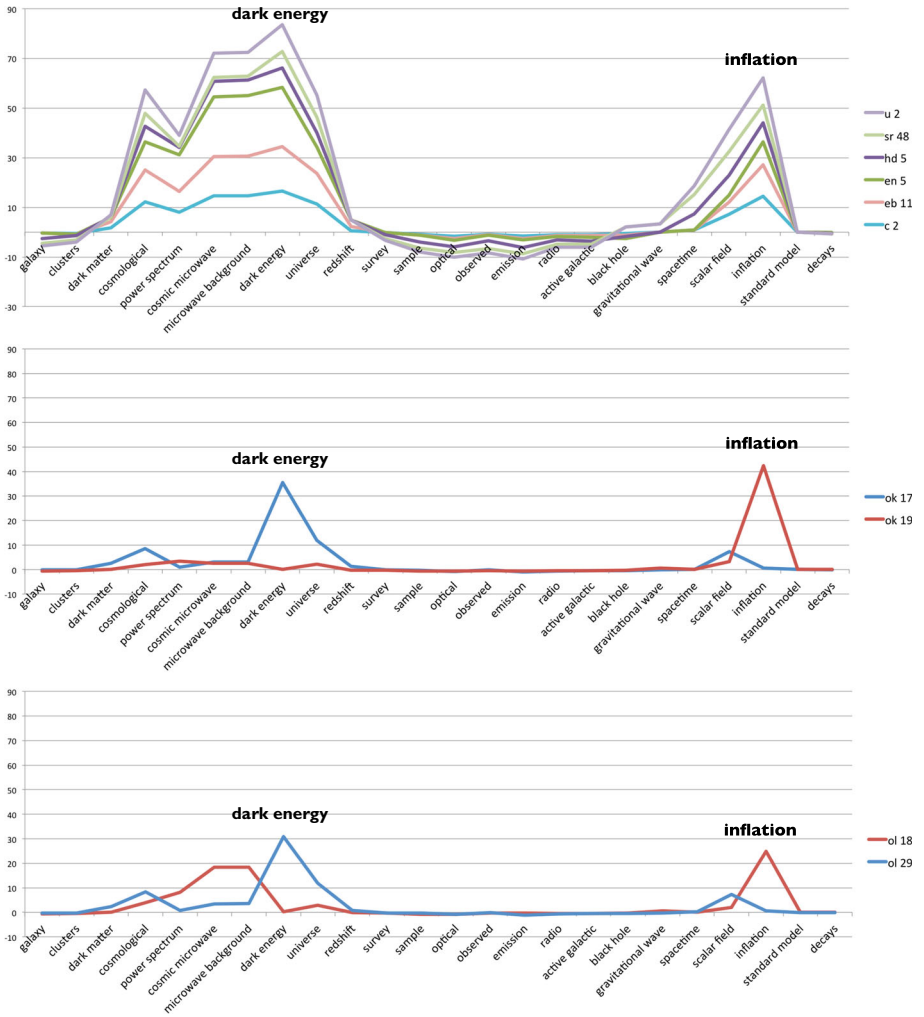


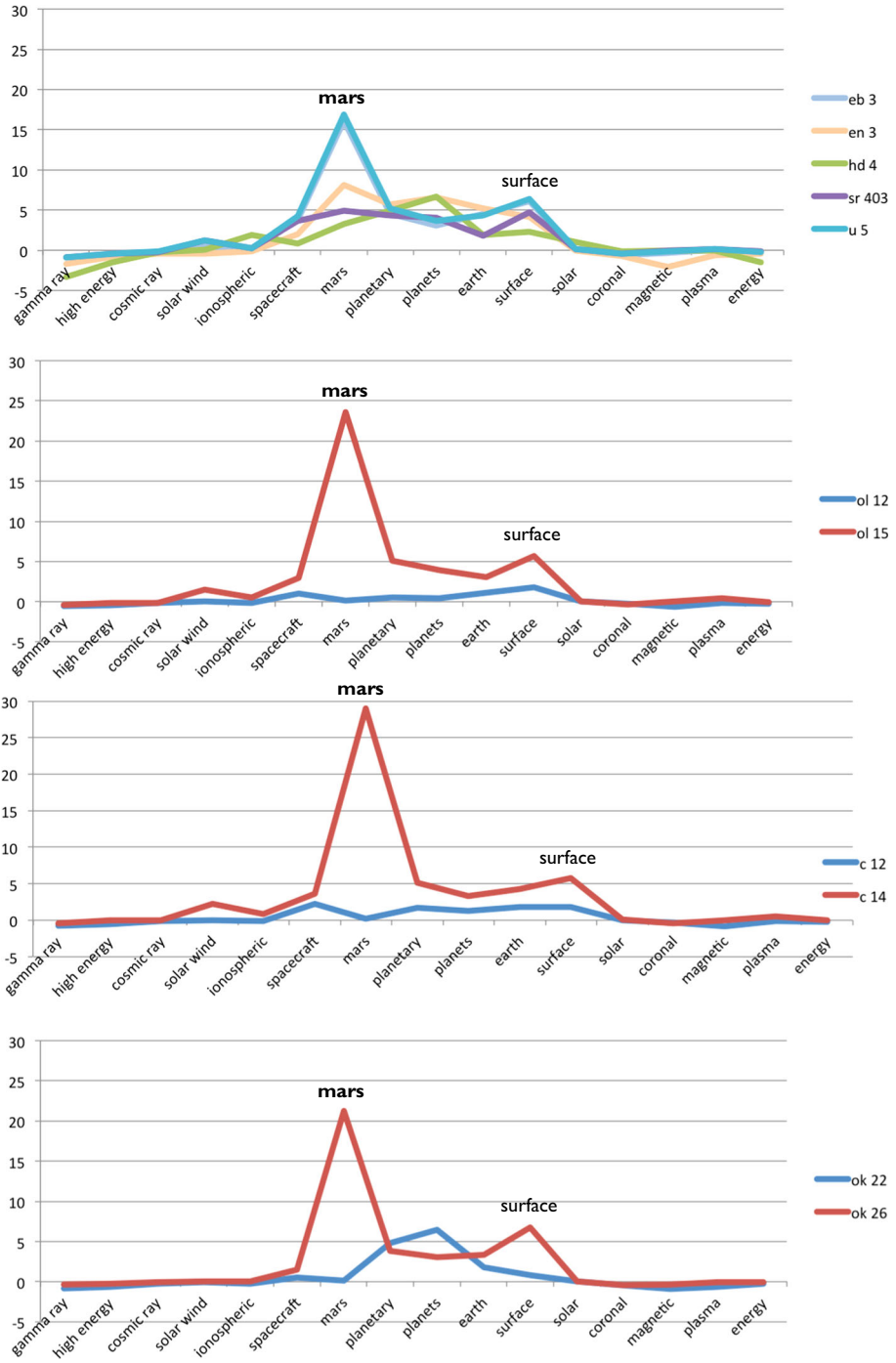

**Fig. 7** Lexical fingerprints of clusters that include the shared data sets 5 and 13. Whereas most topic solutions assign the two shared paper sets to a single topic (*top diagram*), solutions *ol* and *ok* (*below*) distinguish the topics of 'dark energy' (ok17, ol 29) and 'inflation' (ok19, ol18)

## Findings: specific comparisons

In this section we compare pairs of solutions that differ with regard to some specific aspect of their approach (e.g. the same data model was used but different clustering algorithms) and explore whether we can develop hypotheses how differences between solutions link to differences of the approaches.

### Local versus global data

Seven out of the eight approaches represent topics constructed from local data in the sense that they are based exclusively on the information contained in the *Astro Data Set*. By


30
25
20
15
10
5
0
-5
mars
surface
gamma ray
high energy
cosmic ray
solar wind
ionospheric
spacecraft
mars
planetary
planets
earth
surface
solar
coronal
magnetic
plasma
energy
eb 3
en 3
hd 4
sr 403
u 5
ol 12
ol 15
c 12
c 14
ok 22
ok 26

◀**Fig. 8** Lexical fingerprints of the clusters associated with shared data sets 8 and 10. Most topic solutions assign the two shared paper sets to a single topic (top diagram). However, solutions *ol*, *ok*, and *c* (below) assign them to two different topics. One of them is a topic described by the terms 'mars' and 'surface' (ol15, c14). The fingerprint for the second topic (ol12, c12) is less well expressed, because key terms for their characterization such as 'comet' or 'asteroid' are not included in the vocabulary used for the construction of the lexical fingerprint

contrast, *sr* generates topics by mapping the documents of the *Astro Data Set* onto the partitioning of the STS global map of science, the clustered direct citation network of a much larger data set of publications. The underlying data covers a longer time period, 1996–2012, and publications from all areas of science, about 49 million documents in total (Boyack 2017a).

The topical structure that *sr* constructs by embedding the *Astro Data Set* into a global context greatly varies in resolution across the different domains, as can be seen from Figs. 4 and 5. For a detailed analysis see "Details of analysis of local data versus global data" in Appendix. The domains of *Gravitation and Cosmology*, *Astrophysics*, and *Solar Physics* are highly concentrated with almost all documents included in one or two large clusters. In the other domains, documents are dispersed across a larger number of clusters. This suggests that in those domains documents have links to many other parts of the scientific literature outside of the data set. As demonstrated in the companion article on the *sr* solution (Boyack 2017a), many of the smaller topics in the *sr* solution are instances where an astronomy-related application constitutes a part of another, much larger discipline. For example, some of the small topics found by *sr* in *Planetary Science* and *Space Science* seem to have clear links to geology or atmospheric and climate science.

This greater resolution of topics at the periphery of the *Astro Data Set* provides an alternative perspective to the one provided by solutions that construct cohesive clusters in those domains. The appropriateness of either may depend on the purpose of the topic extraction. For example, Boyack (2017a) suggests that a journal based field delineation that neglects the global context of an area of research is increasingly inappropriate to capture topics of research given the increasing interdisciplinarity of research. At the same time, the *sr* solution lacks topical resolution for the two largest domains that constitute the core of the field. The use of an aggregated version of the global science map to create *sr* is likely responsible. As discussed in Boyack (2017a), an aggregated version of the global science map was used so that the number of clusters would correspond more closely with the other solutions submitted to the comparison exercise reported here.[10]

## Citation based versus semantic data models

Another fundamental distinction between approaches is whether their data model uses citation links or lexical similarities in the meta data of an article (such as title and abstract) to relate documents to one another. While citation is a technically unambiguous signal (either there is a citation from one document to another or there is not), there is the potential issue of a social distortion of citation patterns due to rivalries between authors who may avoid citing each others work even though it is related, or bias due to the Matthew's effect in favor of renown authors that attract citations even though other works may be equally relevant but do not get cited as much. By contrast, a semantic approach could be seen as being less vulnerable to such behavioral distortions. However, its signals

[10] It was produced by an algorithmic merging of small topic clusters based on semantic similarities which leads to the construction of large topic clusters, called regions'.

may be technically ambiguous and lead to false positives when the same term is used in different specializations to indicate different concepts. This will occur less often if the data set is focused on a specific scientific area such as the one represented by the *Astro data Set*.

Four of the eight solutions included in our comparison are exclusively based on citation information (*c, u, eb* and *hd*), whereas three have a semantic component in their data model. All of these latter three, however, are some sort of hybrid and none is based purely on semantic information. Solution *sr* is based on a fine-grained clustering of a direct citation network that covers all of science and then uses semantic information to merge clusters again to larger topics. Solution *en* is generated by an explicitly hybrid approach that combines bibliographic coupling and document similarities based on terms extracted using an NLP approach. Finally, solutions *ok* and *ol* are based on what could be termed a 'hyper' semantic data model: it interprets all types of fields in the bibliographic record of a publication as an entity, e.g. author name, article title, journal name, reference. It then constructs a lexical profile for each instance of an entity for all entities by constructing a vector based on the number of publications where those instances co-occur with a given term or subject extracted from the entire data set. To relate articles to one another, for each article the lexical profiles of its entities are combined into one vector. References are one of the entity types included, such that a citation based signal is reintroduced through the back door: two articles that cite the same document will be more similar to each other since the lexical profiles of the respective instance of the reference entity will be the same. The heterogeneity within each of the two sets of solutions, citation based versus (hybrid) semantic, with regard to resolution (number of clusters), clustering algorithms used, and data models, is so great that we restrict a detailed analysis to subsets that reduce this heterogeneity.

*Direct citation* (*c*, *u*) *versus hypersemantic data model* (*ol*, *ok*)

Based on the NMI calculations, these four solutions form a core group of very similar solutions (see Fig. 3), even though they are based on two very different inputs to their data models. The similarity between these two sets of solutions is also reflected in the affinity networks in Figs. 4 and 5.

One significant difference between the two sets of solutions, however, is not captured by the NMI scores because their calculation is based only on those documents that are included in both solutions that are being compared. The direct citation based solutions and the hypersemantic solutions differ substantially in coverage. Whereas the hypersemantic data model includes all documents in the *Astro Data Set*, the direct citation based approach is applicable only to documents that have direct citations to other documents in the data set, and solutions *c* and *u* specifically included only the giant component of the direct citation network.

Interestingly, we find that a large proportion of the ca. 9000 documents omitted from the citation based solutions contribute to a single large topic in solution *ok* (ca. 6300 documents), and in solution *ol* (ca. 7800 documents). Based on the cluster labels the topic seems to be space missions (see "Details of analysis of citation based versus semantic data models" in Appendix for details). We observe that these two clusters exhibit the lowest within cluster citation rate[11] ($\leq 20\%$), much lower than the within citation rates in the majority of clusters ($\geq 40\%$). The resulting sparsely connected direct citation network makes it less likely for a citation based approach to construct a topic out of these documents.

[11] Calculated as the average percentage of references per article that point to journals not included in our data set of 59 journals.

*Bibliographic coupling* (*eb*) *versus hybrid* (*en*)

As reported above, the purely biographic coupling solution *eb* and the hybrid solution *en* do not expose a great similarity based on their NMI score, although they partially overlap in their data model (a bibliographic coupling network), use the same clustering algorithm (Louvain), and have a similar number of clusters. A first observation from the affinity networks in Fig. 9 is that the addition of a lexical component in the data model for *en* has led to a greater aggregation of documents into topics: although *en* covers a slightly larger number of documents (109,376 versus 108,512), it distinguishes only 11 topics while *eb* distinguishes 13 topics.

A more detailed analysis of the differences in topological structure of the affinity networks in Fig. 9 is documented in the "Details of analysis of citation based versus semantic data models" in Appendix. It suggests that some of the distinctive features of solution *en* when compared to *eb* can probably be explained by aggregation effects due to the lexical component in the data model of *en*, such as the relatively larger sizes of the *Planetary Science* and *Astroparticle Physics* domains when compared to *eb*. We find in our analysis that in *eb* research on extra-solar planets seems to be included primarily in *Astrophysics*, whereas in *en* it is split between *Planetary Science* and *Astrophysics*, thereby contributing to a bigger size of *Planetary Science* in *en*. The search for extra-solar planets is to a large extent about the close observation of stars and variations in their movement or radiation. Hence we can expect publications on the search for extra solar planets to frequently reference literature on stellar observations, resulting in close ties in the citation based data model of *eb*. This connection is weakened in *en* because of its data model. It considers also lexical similarity, such that the use of terms relating to 'planets' in the publications about extra-solar planets strengthens their links into the planetary science literature. We speculate that a similar effect may be at work with regard to literature on supergravity, resulting in *en* in a greater aggregation of documents into the *Astroparticle Physics* domain (see "Details of analysis of citation based versus semantic data models" in Appendix for details).

Second, we notice that in solution *en* the topic representing the *Solar Physics* domain is curiously placed at the *Gravitation and Cosmology* end of the affinity network in contrast to the affinity networks of the other solutions that place it at the other end, alongside *Planetary Science*. We find in a search of titles of documents that in *en* the term 'plasma' is relatively concentrated with 71% of occurrences in the single *Solar Physics* topic, whereas in *eb* the concentration of the term 'plasma' in the *Solar Physics* topic is considerably lower, with 52% of occurrences. Further, the affinity network of *en* in Fig. 9 shows that the single document cluster that constitutes *Solar Physics* has relatively strong citation links to the topics of specific types of radiation sources ('gamma-ray sources', 'x-ray sources', and 'gravitational wave sources'). This suggests that due to the lexical component in the data model of *en*, documents that could have been placed into those latter topics based on citations were subsumed instead into the solar physics topic because of their use of terms like 'plasma'.

These observed effects of a lexical component in the data model on the topics constructed raise questions about a conceptual shift in perspective on topical relatedness and on what constitutes a topic. From a theoretical standpoint, citations constitute part of the scientific discourse and according to Gläser (2006) are an important step in the integration of the scientific knowledge base of a research specialty. By contrast, the lexical identity of terms, even when based on agreement about their semantic meaning, does not reflect the same type of topical relatedness as enacted and constructed in scientific discourse. The semantic component in the hybrid approach of *en* constructs topical relatedness based on

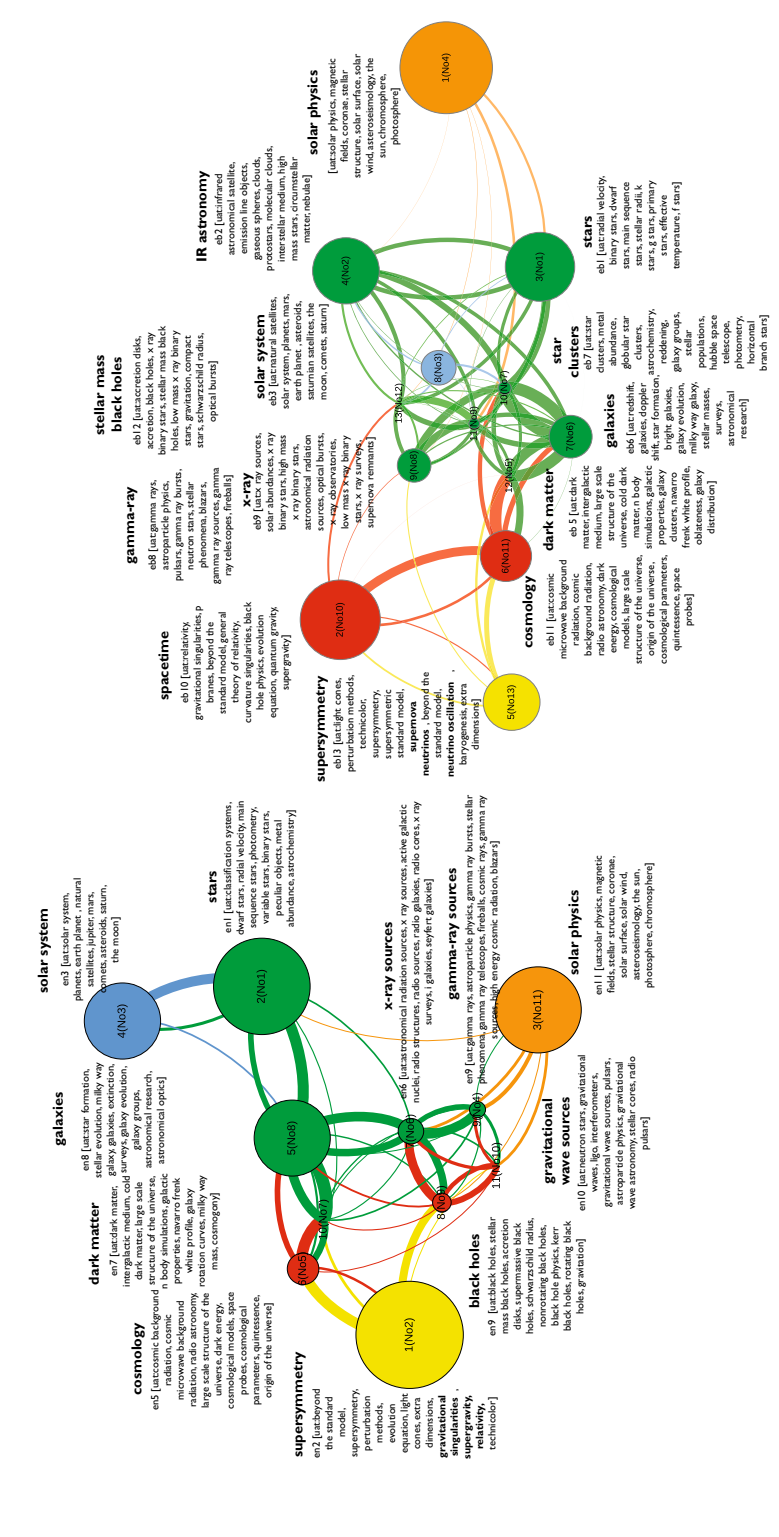



**Fig. 9** Annotated affinity networks for solutions *en* (*left*) and *eb* (*right*). *Labels* used are the entropy-based thesaurus term labels. The labels in ***bold*** are human generated to support readability. *Nodes* are numbered by size, and *numbers in brackets* refer to the original cluster index given by the creators of the solutions (e.g. *1(No4)* refers to the largest cluster in the network, with original cluster index 4)

lexical agreement in order to protect against social distortions in citation patterns. However, it does so at the price of de-emphasizing the discursive context that is expressed in citation patterns (that would make a distinction between plasma in the study of active galaxies versus plasma in the context of solar physics). The notable aggregation effects in the construction of topics by solution *en* discussed in this section suggest that this shift in perspective and its implications for the topical structure it constructs deserve further investigation.

### Clustering: local clustering versus global clustering

Seven out of eight approaches use a global clustering approach. The clustering algorithms they use are designed to take information from the entire network into account when defining document clusters and they produce document clusters that are disjoint, that is each document is assigned to one cluster only. By contrast, the memetic clustering algorithm used to produce the *hd* solution builds clusters locally, starting from seeds and evaluating the immediate environment of each cluster to decide on cluster membership of a node. This approach produces overlapping clusters and assigns to each document a strength of membership. The property of producing overlapping clusters would seem more appropriate to theoretical considerations about the poly-hierarchical nature of topics as argued in Havemann et al. (2017), however it shares with the other approaches the unresolved methodological challenge of evaluating the appropriateness of the topics it constructs.

To explore the difference between the local clustering approach and the global clustering approach we compare solution *hd* to solution *c* that was produced using the same data model (a direct citation network) but with a different clustering algorithm. In our exploratory investigation we pursue the following strategy: we select two domains to investigate in detail, namely *Astroparticle Physics* and *Gravitational Physics and Cosmology*. We compare the lexical fingerprints of the topics that *hd* and *c* constructed in these two domains to see how they differ. We report our findings below (see "Details of analysis of local versus global clustering" in Appendix for details).

When comparing fingerprints of topics in *Astroparticle Physics* for *hd* and *c* (see Fig. 10) we observe that both solutions agree in identifying two major topics, one with a peak at 'qcd' (*c7*: 5363 documents, *hd10*: 5701 documents) and one with a peak at 'standard model' (*c8*: 5211, *hd11*: 5165 documents). In addition, *hd* offers a third distinct topic with a peak at 'decays' (hd18: 1812 documents). All other topics identified by *hd* seem to be variations of these three topics and tend to be smaller.

The comparison of lexical fingerprints for topics in the domain *Gravitational Physics and Cosmology* reveals a similar picture, however without the discovery of a distinct new topic by hd (see "Details of analysis of local versus global clustering" in Appendix): the two solutions agree in the identification of four major topics, and the additional topics that *hd* identifies in the domain *Gravitational Physics and Cosmology* all seem to be smaller variants of the major ones.

This suggests that the local clustering approach reproduces the major topics identified by the global clustering approach. Importantly, it further offers an additional more focused topic ('decays') that is not distinguished in the global clustering solution. Also, it produces at times a scatter of sets of smaller topics that are largely redundant and seem to be close variants of a larger topic within the solution. We further observe with regard to the major topics retrieved that the local approach, since it allows for overlap, produces at times more

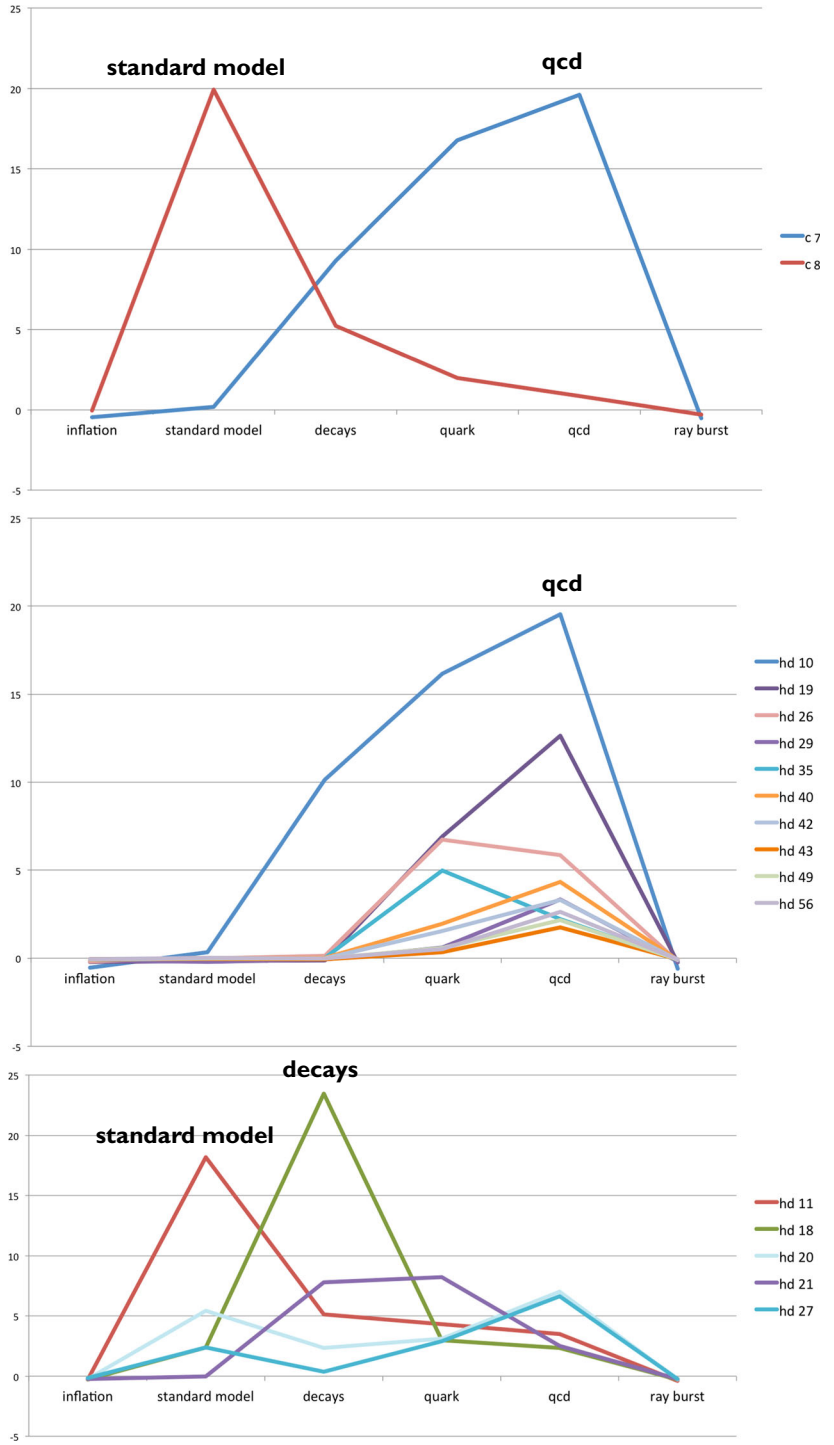



**Fig. 10** Comparison of (partial) fingerprints for *Astroparticle Physics* topics in solutions *hd* and *c*

inclusive topics (see discussion of hd5 versus c2 in "Details of analysis of local versus global clustering" of Appendix).

## Discussion

In this paper we focus on the similarity and dissimilarity of different topic extraction methods and the topic solutions that they deliver. On a general level, as for instance relevant to information retrieval, we found that there is a big overlap in the representations delivered by the different approaches, especially if one views the topical structure of a field as a continuum (a cognitive landscape) rather than a discrete categorization, and focuses less on the rather artificially drawn borders of topics and more on their relative distance, as visualized e.g. by topic affinity networks.[12] However, to study the emergence of something new (starting at a microscopic level), and for evaluation when applied to a micro level such as groups or institutions, small differences between the topic structures constructed by these approaches matter.

The comparison of approaches in this paper provides first insights into the variability of the solutions delivered, and first suggestions of how specific features of approaches shape the topical structures that they construct. For a detailed analysis of how choices of data models, clustering algorithms and parameter values link to specific features of solutions, a more systematic and comprehensive experimental design is needed that varies only one variable at a time. One would ideally study the complete space of solutions generated by all combinations of data models and clustering algorithms and a systematic scanning of parameters that determine the resolution of a solution. This would allow us to evaluate the relative role of data model and algorithms in producing similarities and differences, and to explore the possible influence of the number of clusters we define by the various resolution parameters of the different algorithms. If we had similar numbers of clusters, and each algorithm is forced to distribute the papers between this number of clusters: How likely would it be that the clusters will be similar? Under what conditions could they be different? In terms of the effort required a study of the entire solution space has been outside the scope of this activity. Instead, one of the main contributions of this paper (and its companion papers) are methods for investigating differences between solutions, such as Ariadne, the lexical fingerprint analysis, and the affinity network visualization that we expect will be valuable in such a future undertaking. One realization is that we still lack tools and methods to compare clustering solutions that generate overlapping topics.

The specific observations that we made in the comparative analysis of solutions give rise to a number of questions: The first observation concerns the similarity of solutions based on the 'hypersemantic' data model (*ol* and *ok*) to solutions based on a direct citation network (*c* and *u*); this was a rather surprising finding. Does it suggest a relative robustness of the topical features that are exposed by these methods?

Further, it has been interesting to see in our preliminary analysis that the local clustering approach *hd* that allows for overlapping topics to form, not only reproduces the major topics constructed by the other approaches (along with a scatter of smaller, similar topics), but also some new topics, not detected by the other approaches. Does this suggest a greater

[12] An anonymous reviewer of the introduction article of this special issue raised the question whether topic extraction solutions would not look much more similar to each other if one viewed topics as a continuous cognitive landscape instead of as a discontinuous categorization. Indeed, this resonates very well with our observations from the actual empirical comparison of topic extraction solutions in this article.

sensitivity of this method do detect 'bridging or 'emerging' topics that tend to be suppressed by approaches that only allow for disjoint topics?

Finally, the peculiarities of the *en* solution compared to the other (disjoint) solutions in this comparison might be due to the fact that it reflects a different perspective on the data. But we do not yet have a good grasp on how to identify and characterize such alternate perspectives on topical structures, and it remains an open challenge to establish the validity and usefulness of the different perspectives. We have encountered this issue in our discussion of external versus internal perspectives (see "Local versus global data" section), as well in our discussion of the hybrid lexical approach in contrast to a citation based approach (see "Citation based versus semantic data models" section). We lack well articulated links between (alternate) theories of what constitutes a scientific topic and the operationalization of topics in topic extraction approaches through the way the data is modeled and the clustering algorithm is designed [see discussion in the introduction to this special issue (Gläser et al. 2017)]. We envision as a next step in order to move toward a theory of topical structures in scientific fields, to take a set of empirically extracted topical structures such as the ones in this article and explore the different uses and properties of those perspectives in interaction with topic experts and users of topic extraction results (such as science policy consultants).

## Conclusions

It seems evident that uncertainty about the appropriateness of a topical structure constructed by a topic extraction approach cannot be removed. Uncertainty may relate to the accuracy and completeness of the raw data used, to the validity of the operationalization of topics by the choice of data model and clustering algorithm, to the existence of undetected coding bugs, as well as to the interpretation of the topic extraction outcome [see conceptualization of uncertainty from Arthur Petersen's work on climate modeling (Petersen 2012)]. To which extent such uncertainty is acceptable would seem to depend on the purpose of the topic identification; it makes a difference whether the identification of topics is done to contribute a metric analysis to a science history argument (Burger and Bujdosó 1985), to be used during consultation in the discourse with experts, or for evaluation purposes (Hicks et al. 2015).

We would like to encourage future work on topic identification to re-use part of the framework developed here to better describe and distinguish approaches (raw data, data model, algorithm, parameters). Ideally in times of open science, algorithms and raw data should be shared—using existing Trusted Digital Repositories, which allow to find software and data also in the long term (Dillo et al. 2013). Because this proves to be problematic due to mixed ownership of data products used we would like to call the community—probably also in collaboration with the private information services—to create benchmark datasets, which can be shared openly. The lack of such benchmark datasets [already remarked on by the IR community (Mayr and Scharnhorst 2015)] seems to hamper the further methodological development in the field of scientometrics, and unnecessarily restrains discussions as conducted in this special issue.

The general lack of benchmark data sets also widens the gap between those operating in the field of bibliometrics as professionals in research evaluation (bibliometrics as service), those applying bibliometrics occasionally for such purposes and those applying bibliometrics as one method next to others to better understand the dynamics of science. If

exclusive access to specific databases and tacit knowledge on the implementation of certain algorithms becomes the dominant regime, the further development of bibliometric methods comes to a stop, and it becomes more probable that other communities will re-enter into the same problem space, by simply ignoring lessons learned in the history of bibliometrics.

As a first step, the group behind this special issue on "Same Data, Different Results" reached an agreement with the primary owner of the *Astro Data Set*, originally Thomson Reuters, now Clarivate Analytics, to enable us to share the *Astro Data Set* with the wider scientific community. We would like to invite you to join us in constructing topical structures from this data set and in comparing our approaches and results. See the call for participation in a topic extraction challenge in this special issue (Boyack et al. 2017) or the website www.topic-challenge.info for further information.

**Acknowledgements** We gratefully acknowledge our colleagues Nees Jan van Eck, Wolfgang Glänzel, Frank Havemann, Michael Heinz, Bart Thijs for contributing by providing topic extraction solutions and for lively discussions of the topic extraction exercise at the series of workshops held from 2013 to 2015 in Berlin and Amsterdam. We further thank Michael Heinz for providing NMI calculations and shared document set data for the comparative analysis.

**Funding** Part of this work has been funded by the COST Action TD1210 Knowescape, providing funds for meetings, and mutual visits; and the EC funded project ImpactEV. Theresa Velden would like to thank Carl Lagoze for generous support of her work on this project.

## Appendix

### Data files

The following data files (doi:http://dx.doi.org/10.17026/dans-zzq-z4xh) are made available with the publication of this article through the easy online data archive, operated by Data Archiving and Networked Services (DANS) at https://easy.dans.knaw.nl/:

1. Data file with lexical fingerprint scores for all cluster (≥100 documents) in all solutions.
2. Data file with entropy based word labels for all clusters (≥100 documents) in all solutions.
3. Data file with entropy based thesaurus labels for all clusters (≥100 documents) in all solutions.
4. Data file with journal signature (for up to largest 35 clusters in all seven disjoint solutions)
5. Affinity network files for all seven disjoint solutions (gephi and gefx formats).

### Data set: journal titles

ACTA ASTRONOM, ADV SPACE RES, ANN GEOPHYS, ANNU REV ASTRON ASTROPHYS, ANNU REV EARTH PLANET SCI, ASTROBIOLOGY, ASTRON ASTROPHYS, ASTRON ASTROPHYS REV, ASTRON GEOPHYS, ASTRON J, ASTRON LETT, ASTRON NACHR,ASTRON REP, ASTROPART PHYSICS, ASTROPHYS BULL, ASTROPHYS J, ASTROPHYS J LETT, ASTROPHYS J SUPPL SER , ASTROPHYS SPACE SCI, ASTROPHYSICS, BALT ASTRON, BULL ASTRON SOC INDIA, C R PHYS, CELEST MECH DYNAM ASTRON, CHIN ASTRON ASTROPHYS-ENGL TR, CHINESE J ASTRON ASTROPHYS, CLASS QUANTUM

GRAVITY, CONTRIB ASTRON OBS S, COSM RES, EARTH MOON PLANET, EXP ASTRON, GEN RELATIV GRAVIT, GEOPHYS ASTROPHYS FLUID DYNAM, GRAVIT COSMOL, GRAVIT COSMOL-RUSSIA, IAU SYMP, ICARUS, INT J ASTROBIOL, INT J MOD PHYS D, J ASTROPHYS ASTRON, J COSMOL ASTROPART PHYS, J KOREAN ASTRON SOC, JBIS-J BR INTERPLANET SOC, KINEMAT PHYS CELEST+, MON NOTIC ROY ASTRON SOC, NEW ASTRON, NEW ASTRON REV NUOVO CIMENTO C-GEOPHYS SPACE, OBSERVATORY, PHYS REV D, PLANET SPACE SCI, PUBL ASTRON SOC AUSTRALIA, PUBL ASTRON SOC JPN, PUBL ASTRON SOC PAC, RES ASTRON ASTROPHYS, REV MEX ASTRON ASTROFIS, SOL PHYS, SOLAR SYST RES, SPACE SCI REV.

### Comparison metric: Normalized Mutual Information

We consider two clusterings $\mathbf{C} = \{C_1, C_2, \ldots, C_m\}$ and $\mathbf{D} = \{D_1, D_2, \ldots, D_n\}$ of the document set $M$. To compare the two clusterings we calculate the matrix of the sizes of intersections of clusters from $\mathbf{C}$ and $\mathbf{D}$:

$$\mathbf{DC} = \begin{array}{c} \\ D_1 \\ D_2 \\ \vdots \\ D_n \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \ldots & C_m \end{array} \\ \begin{pmatrix} dc_{1,1} & dc_{1,1} & \ldots & dc_{1,m} \\ dc_{2,1} & dc_{2,2} & \ldots & dc_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ dc_{n,1} & dc_{n,2} & \ldots & dc_{n,m} \end{pmatrix} \end{array}$$

where $dc_{i,j} = |D_i \cap C_j|$ is the number of elements in the intersection of the two clusters $D_i$ and $C_j$. For the rows and columns we introduce the abbreviations:

$$dc_{i,*} = \sum_{j=1}^{m} dc_{i,j}$$

$$dc_{*,j} = \sum_{i=1}^{n} dc_{i,j}$$

The number of elements (documents) in M we denote by $N$. Hence:

$$N = \sum_{i=1}^{n} dc_{i,*} = \sum_{j=1}^{m} dc_{*,j} = \sum_{i=1}^{n} \sum_{j=1}^{m} dc_{i,j} = |M|$$

Mutual information is formally defined as:

$$\mathrm{MI}(C, D) := \sum_{i=1}^{n} \sum_{j=1}^{m} P(D_i, C_j) \cdot \log\left(\frac{P(D_i, C_j)}{P(D_i) \cdot P(C_j)}\right)$$

where $P(D_i, C_j)$ denotes the probability that a document is in cluster $D_i$ and in cluster $C_j$, $P(D_i)$ denotes the probability that a document is in cluster $D_i$, and $P(C_j)$ denotes the probability that a document is in cluster $C_j$. If we estimate these probabilities by dividing the number of observed events, $dc_{i,j}$, $dc_{i,*}$, $dc_{*,j}$, respectively, with the total number of documents N, the Mutual Information can be rewritten as:

$$\mathrm{MI}(C,D) = \frac{1}{N} \cdot \sum_{i=1}^{n} \sum_{j=1}^{m} dc_{i,j} \cdot \log\left(\frac{N \cdot dc_{i,j}}{dc_{i,*} \cdot dc_{*j}}\right)$$

We use Normalized Mutual Information (*NMI*) in our comparisons, which is defined as follows:

$$\mathrm{NMI}(C,D) := \frac{MI(C,D)}{H(C,D)}$$

where the joint entropy $H(C, D)$ is defined as:

$$H(C,D) := -\sum_{i=1}^{n} \sum_{j=1}^{m} P(D_i, C_j) \cdot \log(P(D_i, C_j))$$

and can be rewritten as:

$$H(C,D) = \frac{1}{N} \cdot \sum_{i=1}^{n} \sum_{j=1}^{m} dc_{i,j} \cdot \log\left(\frac{N}{dc_{i,j}}\right).$$

The Mutual Information MI is always smaller or equal to the (information) entropy H: $MI(C,D) \leq H(C,D)$. If the two clusterings are identical, than $\mathrm{MI}(C,D)$ and $H(C, D)$ are equal and the highest value of $\mathrm{NMI} = 1$ is attained.

## Cluster-level labels

In Table 4 we provide for all eight solutions word based cluster labels for clusters $\geq 100$ documents.

**Table 4** Word labels for clusters

| No. | Size | Word labels |
|---|---|---|
| c1 | 14,873 | Galaxies, redshift, star formation, sample, active galactic, agn, gas, galaxy clusters, digital sky, sloan digital |
| c10 | 3904 | Black holes, spacetimes, horizon, ads, solutions, metric, dimensional, supergravity, static, spherically symmetric |
| c11 | 3527 | Pulsar, supernova remnant, psr, snr, neutron star, wind nebula, anomalous x, remnant snr, radio pulsars, magnetar |
| c12 | 3413 | Asteroid, comet, main belt, kuiper belt, meteor, perihelion, bodies, solar system, albedo, trans neptunian |
| c13 | 3392 | Eclipsing binary, star, asteroseismic, chemically peculiar, wilson devinney, delta scuti, eta carinae, contact binary, hd, pulsation |
| c14 | 3355 | Mars, titan, atmosphere, water, deposits, cassini, mars express, ice, venus, moon |
| c15 | 3182 | Grb, ray bursts, gamma ray, afterglow, bursts grbs, sn, explosion, swift, type ia, supernova sn |
| c16 | 3156 | Gravitational wave, lisa, inspiral, ligo, wave detectors, laser interferometer, binary black, waveforms, numerical relativity, post newtonian |
| c17 | 2625 | Blazar, bl lac, jet, lac objects, radio galaxies, synchrotron, ultra high, radio, radio sources, 3c |

**Table 4** continued

| No. | Size | Word labels |
|---|---|---|
| c18 | 2228 | Ionospheric, auroral, radar, substorm, geomagnetic, magnetopause, iri, tec, midnight, field aligned |
| c19 | 2088 | White dwarf, nova, cataclysmic variable, dwarf nova, subdwarf b, wd, sdb, orbital period, sdb stars, superhumps |
| c2 | 8954 | Dark energy, microwave background, cosmic microwave, inflation, cosmological, universe, cmb, power spectrum, background cmb, scalar field |
| c20 | 1963 | Quantum gravity, loop quantum, noncommutative, quantum, quantization, quantum cosmology, algebra, spin foam, casimir, hilbert space |
| c21 | 1839 | Planetary nebulae, pne, post agb, central star, asymptotic giant, nebulae pne, symbiotic, pn, mira, agb stars |
| c22 | 794 | Teleparallel, nutation, lense thirring, lageos, laser ranging, gravitomagnetic, celestial reference, pioneer, gravitational field, grace |
| c3 | 7998 | Solar, coronal, active region, cme, flare, magnetic field, sunspot, mass ejections, quiet sun, chromosphere |
| c4 | 7483 | Planets, brown dwarfs, planet formation, transit, extrasolar planets, star, tauri stars, jup, giant planet, hd |
| c5 | 5704 | Molecular cloud, protostellar, cloud, interstellar, young stellar, star forming, molecules, massive star, forming region, stellar objects |
| c6 | 5597 | Globular clusters, fe h, metal poor, giant branch, stars, red giant, metallicity, galactic globular, horizontal branch, milky way |
| c7 | 5363 | Qcd, quark, meson, lattice, decays, chiral, pi pi, gluon, pion, j psi |
| c8 | 5211 | Standard model, neutrino, higgs, lhc, minimal supersymmetric, lepton, supersymmetric standard, gev, top quark, hadron collider |
| c9 | 5179 | X ray, ray binary, black hole, accretion disk, hard state, ray timing, neutron star, rossi x, timing explorer, xmm newton |
| eb1 | 10,666 | Star, radial velocity, planet, orbital period, hd, transit, binary, eclipsing binary, main sequence, white dwarf |
| eb10 | 11,638 | Spacetime, brane, metric, black hole, quantum, gravitational wave, quantum gravity, solutions, einstein, general relativity |
| eb11 | 9118 | Dark energy, microwave background, cosmic microwave, cosmological, inflation, universe, cmb, power spectrum, background cmb, wmap |
| eb12 | 4639 | Accretion disk, black hole, disk, magnetorotational instability, ray binaries, periodic oscillations, viscous, hard state, migration, angular momentum |
| eb13 | 9538 | Quark, qcd, decays, standard model, flavor, meson, lattice, gluon, bar, leading order |
| eb2 | 10,408 | Molecular cloud, young stellar, dust, protostellar, star forming, cloud, forming region, molecules, iras, brown dwarfs |
| eb3 | 7650 | Mars, asteroid, titan, comet, water, surface, moon, saturn, cassini, ice |
| eb4 | 12,678 | Solar, magnetic field, coronal mass, solar activity, plasma, active region, ionospheric, cme, flare, sunspot |
| eb5 | 4947 | Dark matter, halo, n body, body simulations, cold dark, matter halo, galaxy clusters, navarro frenk, galaxies, frenk white |
| eb6 | 8366 | Galaxies, star formation, sloan digital, digital sky, redshift, sample, sky survey, active galactic, sdss, agn |
| eb7 | 5755 | Globular cluster, photometry, color magnitude, galactic globular, fe h, ngc, metallicity, red giant, reddening, horizontal branch |
| eb8 | 7560 | Gamma ray, pulsar, ray bursts, grb, bursts grbs, high energy, jet, radio, psr, synchrotron |

**Table 4** continued

| No. | Size | Word labels |
|---|---|---|
| eb9 | 5549 | Xmm newton, x ray, kev, chandra, 10 kev, newton observations, ray spectrum, ray emission, cm 2, suzaku |
| en1 | 16,142 | Stars, main sequence, radial velocity, giant branch, photometry, fe h, binary, red giant, asymptotic giant, mass loss |
| en10 | 3662 | Gravitational wave, neutron star, pulsar, wave detectors, ligo, lisa, interferometric gravitational, isolated neutron, radio pulsars, laser interferometer |
| en11 | 14,830 | Magnetic field, solar wind, plasma, coronal mass, ionospheric, active region, waves, solar activity, field lines, reconnection |
| en2 | 17,568 | Qcd, quark, standard model, decays, flavor, gauge, meson, higgs, field theory, symmetry |
| en3 | 13,513 | Mars, solar system, planet, water, asteroid, earth, ice, comet, bodies, surface |
| en4 | 5720 | Gamma ray, grb, ray bursts, cosmic ray, high energy, bursts grbs, afterglow, swift, tev, tev gamma |
| en5 | 7752 | Microwave background, cosmic microwave, dark energy, power spectrum, cmb, background cmb, cosmological, universe, type ia, inflation |
| en6 | 6936 | Xmm newton, x ray, chandra, kev, radio sources, radio, active galactic, ray emission, newton observations, 10 kev |
| en7 | 3713 | Dark matter, cold dark, matter halo, halo, n body, body simulations, wimp, navarro frenk, matter particles, weakly interacting |
| en8 | 13,485 | Star formation, galaxies, formation rate, digital sky, sloan digital, sky survey, molecular gas, gas, sample, h ii |
| en9 | 6055 | Black hole, supermassive black, hole mass, schwarzschild black, horizon, bh, rotating black, kerr black, binary black, hole binaries |
| hd1 | 67,716 | Galaxies, black hole, redshift, cosmological, dark matter, universe, gamma ray, x ray, active galactic, scalar |
| hd10 | 5256 | Qcd, quark, meson, lattice, decays, chiral, pi pi, gluon, pion, j psi |
| hd11 | 4055 | Standard model, lhc, higgs, lepton, top quark, minimal supersymmetric, neutrino, hadron collider, supersymmetric standard, electroweak |
| hd12 | 2878 | Gravitational wave, lisa, wave detectors, post newtonian, inspiral, ligo, numerical relativity, waveforms, binary black, interferometric gravitational |
| hd13 | 5127 | Mars, ionospheric, titan, auroral, altitude, radar, degrees n, summer, spacecraft, magnetosphere |
| hd14 | 3049 | Ads, black holes, spacetimes, hole solutions, supergravity, horizon, quasinormal modes, five dimensional, yang mills, metric |
| hd15 | 3116 | Mars, titan, water, deposits, atmosphere, cassini, mars express, ice, volcanic, venus |
| hd16 | 4076 | Mars, ionospheric, auroral, radar, summer, magnetosphere, spacecraft, magnetopause, dayside, degrees n |
| hd17 | 1345 | Loop quantum, quantum gravity, noncommutative, quantum cosmology, quantization, quantum, spin foam, algebra, hilbert space, immirzi parameter |
| hd18 | 1343 | Decays, b b, branching fractions, pi pi, babar detector, 0 pi, bar 0, k pi, b meson, pep ii |
| hd19 | 1010 | Chemical potential, nambu jona, jona lasinio, finite temperature, lasinio model, polyakov loop, qcd, phase diagram, quark matter, lattice |
| hd2 | 79,344 | |
| hd20 | 1406 | Top quark, parton, fermilab tevatron, leading order, inclusive, higgs boson, cross section, gluon, lhc, tevatron |
| hd21 | 1100 | Meson, j psi, x 3872, pentaquark, qcd sum, charmonium, pi pi, decays, sum rules, charmed |

**Table 4** continued

| No. | Size | Word labels |
|---|---|---|
| hd22 | 1911 | Mars, meteor, deposits, soil, biological, water, microbial, life, martian surface, microorganisms |
| hd23 | 712 | Loop quantum, quantum gravity, quantum cosmology, quantization, spin foam, hilbert space, quantum, immirzi parameter, ashtekar, hamiltonian constraint |
| hd24 | 2019 | Mars, summer, ionospheric, degrees n, winter, deposits, iri, seasonal, water, soil |
| hd25 | 2207 | Ionospheric, auroral, radar, substorm, magnetopause, iri, tec, geomagnetic, midnight, degrees n |
| hd26 | 821 | Lattice qcd, chiral perturbation, chiral, fm, partially quenched, perturbation theory, staggered, quark masses, pion, nucleon |
| hd27 | 949 | fermilab tevatron, parton, 96 tev, gluon, leading order, inclusive, transverse momentum, rapidity, deep inelastic, cross section |
| hd28 | 1201 | Mars, deposits, water, soil, microbial, martian surface, life, microorganisms, biological, martian atmosphere |
| hd29 | 734 | Holographic, ads cft, yang mills, dual, mills theory, super yang, xs, gauge theory, cft correspondence, string |
| hd3 | 22,167 | Galaxies, redshift, active galactic, star formation, agn, galactic nuclei, sample, clusters, quasar, sloan digital |
| hd30 | 914 | Noncommutative, casimir, teleparallel, seiberg witten, ads 5, energy momentum, yang mills, super yang, algebra, pp wave |
| hd31 | 444 | Titan, huygens probe, cassini, haze, methane, descent, photochemical, aerosols, ch4, disr |
| hd32 | 1101 | Auroral, substorm, magnetopause, ionospheric, plasma sheet, cluster spacecraft, magnetosheath, field aligned, superdarn, dayside |
| hd33 | 829 | Mars, deposits, microorganisms, microbial, life, martian surface, crater, rock, water, volcanic |
| hd34 | 1048 | Ionospheric, iri, tec, electron content, total electron, ionosonde, fof2, gps, degrees n, reference ionosphere |
| hd35 | 379 | Parton distributions, semi inclusive, generalized parton, deep inelastic, transverse momentum, sivers, inelastic scattering, transversely polarized, unpolarized, transversity |
| hd36 | 974 | Ionospheric, iri, degrees n, tec, electron content, ionosonde, summer, fof2, total electron, winter |
| hd37 | 554 | Teleparallel, bianchi type, lyra, gravitation, energy momentum, saez, collineations, spacetimes, ballester, cosmological models |
| hd38 | 697 | Radar, mesosphere, substorm, auroral, ionospheric, echoes, superdarn, eiscat, backscatter, hf |
| hd39 | 275 | Mercury, hermean, bepicolombo, exosphere, mariner, lunar, messenger, mare, regolith, clementine |
| hd4 | 33,766 | Star, planet, main sequence, mars, hd, atmosphere, jupiter, abundances, radial velocity, period |
| hd40 | 324 | Rapidity, balitsky, diffractive, deep inelastic, pomeron, kovchegov, inelastic scattering, hera, parton, unintegrated |
| hd41 | 482 | Microorganisms, microbial, life, biological, dose, bacterial, mars, human, bacillus, spores |
| hd42 | 347 | Generalized parton, parton distributions, diffractive, balitsky, pomeron, kovchegov, rapidity, hera, light front, deep inelastic |
| hd43 | 181 | Landau gauge, gluon propagator, gribov, ghost, zwanziger, faddeev popov, dyson schwinger, propagators, yang mills, coulomb |
| hd44 | 389 | Meteor, leonid, ablation, radiant, radar, geminid, video, trails, quadrantid, shower |

**Table 4** continued

| No. | Size | Word labels |
|---|---|---|
| hd45 | 342 | Bianchi type, lyra, spatially homogeneous, collineations, perfect fluid, saez, ballester, spacetimes, cosmological models, vi0 |
| hd46 | 542 | Iri, ionospheric, tec, electron content, ionosonde, fof2, total electron, reference ionosphere, international reference, content tec |
| hd47 | 268 | Noncommutative, seiberg witten, moyal, algebra, nc, deformed, fuzzy, theta, field theories, superspace |
| hd48 | 533 | Ionospheric, tec, electron content, gps, total electron, equatorial, esf, content tec, stations, fof2 |
| hd49 | 254 | Balitsky, pomeron, kovchegov, rapidity, diffractive, hera, deep inelastic, inelastic scattering, kuraev, lipatov |
| hd5 | 18,669 | Scalar field, dark energy, inflation, cosmological constant, gravity, spacetime, microwave background, cosmic microwave, brane, universe |
| hd50 | 437 | Thermospheric, ionospheric, gps, tec, electron content, esf, champ, equatorial, total electron, ionosonde |
| hd51 | 442 | Mesospheric, summer, esf, pmse, degrees n, tec, equatorial, brazil, noctilucent, ionosonde |
| hd52 | 253 | Ads 5, super yang, twistor, superconformal, xs, yang mills, sym, maldacena, mills theory, magnon |
| hd54 | 164 | Dome, dimm, isoplanatic, scidar, seeing, antarctic, wavefront, pedro martir, san pedro, site |
| hd55 | 188 | Casimir, dirichlet, plates, robin, plana, momentum tensor, massless scalar, vacuum polarization, vacuum expectation, conductor |
| hd56 | 183 | Balitsky, kovchegov, pomeron, diffractive, rapidity, hera, deep inelastic, inelastic scattering, kuraev, lipatov |
| hd57 | 277 | Body problem, restricted three, periodic orbits, three body, equilibrium points, photogravitational, sail, collinear, sitnikov, thrust |
| hd58 | 325 | Ionospheric, esf, tec, brazil, equatorial, thermospheric, fof2, electron content, earthquake, ionosonde |
| hd59 | 301 | Magnetosheath, interball, demeter, vlf, whistler, magnetopause, earthquake, chorus, staff, cluster spacecraft |
| hd6 | 13,128 | Gamma ray, x ray, neutron star, ray bursts, grb, pulsar, high energy, jet, swift, bursts grbs |
| hd60 | 222 | Auroral oval, gic, aurora, magnetosphere, geomagnetically, thermospheric, oval, paraboloid, precipitating, fpi |
| hd61 | 109 | Prasad sommerfield, bogomol nyi, nyi prasad, non abelian, world sheet, fayet iliopoulos, bps, orientational, monopoles, hypermultiplets |
| hd62 | 166 | Ads 5, super yang, xs, magnon, mhv, sym, maldacena, cachazo, recursion, yang mills |
| hd63 | 187 | Superenergy, weyl tensor, bel, spacetimes, killing vectors, petrov, causal, chevreton, electrovacuum, kerr schild |
| hd64 | 131 | Lyra, saez, ballester, bianchi type, tensor theory, gravitation, determinate, lrs, bimetric, bulk viscosity |
| hd65 | 145 | Nutation, iau, stiefel, celestial, kustaanheimo, p03, cip, capitaine, iers, chandler |
| hd66 | 225 | Mesosphere, pmse, mesopause, lower thermosphere, lidar, noctilucent, summer, middle atmosphere, nlc, degrees n |
| hd68 | 147 | Superenergy, weyl tensor, bel, killing vectors, chevreton, causal, petrov, spacetimes, collineations, isometries |
| hd7 | 11,008 | Planet, mars, earth, ionospheric, saturn, jupiter, asteroid, comet, radar, titan |
| hd71 | 141 | Lower thermosphere, mesosphere, qp, meteor radar, semidiurnal, tide, middle atmosphere, uars, mesopause, collm |

**Table 4** continued

| No. | Size | Word labels |
|---|---|---|
| hd75 | 110 | Nutation, iau, p03, capitaine, cip, iers, chandler, celestial, souchay, 2000a |
| hd76 | 148 | Life support, plant, wheat, food, bioregenerative, crops, waste, biomass, cultivation, ecological |
| hd78 | 141 | Demeter, vlf, earthquake, whistler, chorus, lightning, elf, epicenter, hiss, magion |
| hd8 | 9547 | Solar, coronal mass, magnetic, active region, cme, sunspot, flare, mass ejections, solar activity, plasma |
| hd83 | 110 | Demeter, earthquake, vlf, lightning, epicenter, elf, whistler, subionospheric, modis, hiss |
| hd9 | 8755 | Mars, ionospheric, comet, asteroid, spacecraft, titan, radar, saturn, earth, cassini |
| ok0 | 2866 | Seyfert 1, active galactic, narrow line, agn, broad line, galactic nuclei, quasars, line seyfert, nuclei agns, emission line |
| ok1 | 2934 | Lens, microlensing, gravitational lens, rotation curve, spiral galaxies, bars, dark matter, barred galaxies, galaxy, pattern speed |
| ok10 | 4070 | Globular clusters, fe h, metal poor, red giant, metallicity, giant branch, horizontal branch, galactic globular, color magnitude, stars |
| ok11 | 3409 | Ray binary, x ray, hard state, ray timing, rossi x, timing explorer, black hole, accretion disk, neutron star, rxte |
| ok12 | 6022 | Galaxies, star formation, formation rate, deep field, redshift, early type, sample, rest frame, starburst, lyman break |
| ok13 | 5556 | Quantum gravity, quantum, loop quantum, spacetime, general relativity, scalar field, gravity, quantum cosmology, metric, quantization |
| ok14 | 5583 | Coronal, active region, solar, flare, magnetic flux, cme, quiet sun, chromosphere, mass ejections, hinode |
| ok15 | 2569 | Cosmic ray, high energy, gamma rays, tev, hess, ultra high, air showers, tev gamma, extensive air, shower |
| ok16 | 1985 | Microwave background, cosmic microwave, background cmb, cmb, microwave anisotropy, anisotropy probe, wilkinson microwave, wmap, power spectrum, probe wmap |
| ok17 | 2465 | Dark energy, quintessence, universe, phantom, f r, cosmological constant, cosmic acceleration, chaplygin gas, modified gravity, accelerated expansion |
| ok18 | 2206 | White dwarf, cataclysmic variables, dwarf nova, nova, wd, mass transfer, orbital period, secondary star, cvs, superhumps |
| ok19 | 1627 | Inflation, slow roll, curvature perturbation, non gaussianity, inflationary models, curvaton, reheating, cosmological perturbations, f nl, primordial |
| ok2 | 5449 | Transit, star, eclipsing binary, radial velocity, hd, planet, corot, photometric, main sequence, type stars |
| ok20 | 2020 | Gravitational wave, inspiral, ligo, lisa, wave detectors, laser interferometer, waveforms, binary black, space antenna, post newtonian |
| ok21 | 1849 | Grb, ray burst, gamma ray, afterglow, bursts grbs, swift, prompt emission, prompt, fireball, batse |
| ok22 | 4103 | Asteroid, comet, body problem, orbits, kuiper belt, main belt, bodies, mean motion, planets, solar system |
| ok23 | 2071 | Planetary nebulae, asymptotic giant, agb stars, post agb, giant branch, pne, branch agb, agb, central star, mira |
| ok24 | 4592 | Molecular cloud, protostellar, cloud, c 13, star forming, h 2, molecules, hco, forming regions, massive star |
| ok25 | 2622 | Ionospheric, winter, summer, degrees n, mesosphere, tec, electron content, iri, ozone, seasonal |

**Table 4** continued

| No. | Size | Word labels |
|---|---|---|
| ok26 | 3593 | Mars, titan, ice, water, deposits, cassini, co2, methane, atmosphere, surface |
| ok27 | 6292 | Performance, scientific, technology, mission, astronomical, development, research, flight, cost, software |
| ok28 | 4903 | Sn, explosion, wolf rayet, type ia, supernova, ejecta, wr, progenitor, eta carinae, lines |
| ok29 | 3176 | Brown dwarfs, tauri stars, pre main, herbig ae, substellar, circumstellar disks, young, main sequence, disks, low mass |
| ok3 | 3389 | Spacetimes, black hole, horizon, asymptotically flat, reissner nordstrom, metric, einstein maxwell, spherically symmetric, hole solutions, schwarzschild |
| ok30 | 2624 | Pulsar, neutron stars, psr, radio pulsars, anomalous x, magnetar, isolated neutron, soft gamma, millisecond pulsars, axp |
| ok4 | 5420 | Solar wind, magnetosphere, interplanetary magnetic, magnetic field, auroral, plasma, magnetopause, ion, substorm, spacecraft |
| ok5 | 3874 | Standard model, higgs, lhc, minimal supersymmetric, supersymmetric standard, neutrino mass, lepton, right handed, hadron collider, electroweak |
| ok6 | 4721 | Quark, qcd, meson, decays, lattice qcd, pi pi, pion, j psi, form factors, chiral |
| ok7 | 4206 | Galaxy clusters, dark matter, haloes, cluster, n body, weak lensing, intracluster medium, halo mass, 1 mpc, galaxies |
| ok8 | 2195 | Blazar, bl lac, jet, radio sources, lac objects, radio galaxies, synchrotron, radio, flat spectrum, 3c |
| ok9 | 3225 | Yang mills, gauge theory, mills theory, string, supergravity, noncommutative, field theory, supersymmetric, dual, branes |
| ol0 | 3630 | Brown dwarfs, tauri stars, pre main, herbig ae, main sequence, substellar, young, spectral type, circumstellar disks, stars |
| ol1 | 6668 | fe h, globular clusters, metal poor, metallicity, stars, giant branch, red giant, milky way, color magnitude, dwarf spheroidal |
| ol10 | 2252 | Dynamo, solar cycle, helioseismology, sunspot, convection zone, solar activity, p modes, differential rotation, tachocline, cycle |
| ol11 | 6893 | Galaxies, star formation, early type, formation rate, starburst, sample, high redshift, rest frame, stellar populations, surface brightness |
| ol12 | 2450 | Comet, asteroid, main belt, meteor, kuiper belt, perihelion, trans neptunian, belt objects, near earth, albedo |
| ol13 | 9646 | Spacetime, brane, metric, quantum, solutions, horizon, four dimensional, gravity, ads, black hole |
| ol14 | 1539 | Intergalactic medium, reionization, ly alpha, medium igm, absorbers, igm, dlas, damped ly, alpha forest, 21 cm |
| ol15 | 3450 | Mars, titan, cassini, saturn, deposits, ice, atmosphere, mars express, water, mars global |
| ol16 | 7798 | Performance, human, research, scientific, development, technology, earth, mission, astronomical, space |
| ol17 | 2702 | Radio galaxies, blazar, bl lac, radio sources, jet, radio, lac objects, 3c, synchrotron, steep spectrum |
| ol18 | 3412 | Inflation, cosmic microwave, microwave background, non gaussianity, cmb, background cmb, slow roll, wmap, microwave anisotropy, wilkinson microwave |
| ol19 | 2248 | Weak lensing, lens, gravitational lens, strong lensing, clustering, 1 mpc, correlation function, cosmic shear, bias, dark matter |
| ol2 | 5484 | Molecular cloud, protostellar, cloud, interstellar, molecules, young stellar, star forming, h 2, massive star, gas phase |

**Table 4** continued

| No. | Size | Word labels |
|---|---|---|
| ol20 | 5024 | Standard model, higgs, lhc, neutrino, minimal supersymmetric, supersymmetric standard, lepton, top quark, electroweak, hadron collider |
| ol21 | 2709 | Ionospheric, auroral, substorm, radar, magnetopause, geomagnetic, field aligned, plasma sheet, iri, electron content |
| ol22 | 1906 | Cosmic ray, supernova remnant, snr, ultra high, air showers, remnants snrs, extensive air, shower, uhecr, shock acceleration |
| ol23 | 2123 | White dwarf, nova, cataclysmic variables, dwarf nova, subdwarf b, wd, outburst, orbital period, sdb stars, sdb |
| ol24 | 3193 | Seyfert 1, active galactic, galactic nuclei, agn, broad line, narrow line, quasars, hole mass, nuclei agns, line seyfert |
| ol25 | 6095 | Quark, qcd, meson, lattice, decays, pi pi, gluon, chiral, pion, j psi |
| ol26 | 1935 | Galaxy clusters, intracluster medium, sunyaev zel, icm, cluster, zel dovich, medium icm, cooling flow, sz, dovich effect |
| ol27 | 1858 | Gravitational wave, lisa, ligo, wave detectors, inspiral, laser interferometer, binary black, post newtonian, numerical relativity, interferometric gravitational |
| ol28 | 639 | Mond, modified newtonian, newtonian dynamics, dynamics mond, lorentz violation, special relativity, lorentz symmetry, cpt, aether, teves |
| ol29 | 2764 | Dark energy, universe, quintessence, phantom, f r, cosmological models, scalar field, cosmic acceleration, chaplygin gas, equation |
| ol3 | 4198 | Ray binary, x ray, black hole, neutron star, hard state, ray timing, rossi x, timing explorer, accretion disk, ultraluminous x |
| ol31 | 353 | r matrix, collision strengths, impact excitation, electron impact, breit pauli, dielectronic recombination, atomic data, 3p, oscillator strengths, 3s |
| ol4 | 2252 | Pulsar, psr, radio pulsars, anomalous x, neutron star, magnetar, wind nebula, isolated neutron, soft gamma, millisecond pulsars |
| ol5 | 2089 | Planetary nebulae, pne, post agb, asymptotic giant, mira, agb stars, central star, nebulae pne, symbiotic, mass loss |
| ol6 | 3324 | Eclipsing binary, star, wilson devinney, double lined, wolf rayet, hd, contact binary, chemically peculiar, delta scuti, eta carinae |
| ol7 | 3416 | Planet, transit, body problem, extrasolar planets, giant planets, migration, eccentricity, planet formation, three body, restricted three |
| ol8 | 6585 | Coronal mass, solar, cme, active region, flare, mass ejections, magnetic field, magnetic reconnection, chromosphere, transition region |
| ol9 | 2895 | Grb, ray bursts, gamma ray, afterglow, bursts grbs, sn, type ia, swift, explosion, ia supernovae |
| sh1 | 4162 | Solar, magnetic field, coronal, flare, plasma, active region, cme, mass ejections, reconnection, euv |
| sh10 | 1083 | Titan, cassini, atmosphere, mars, saturn, methane, haze, aerosol, ice, surface |
| sh11 | 1066 | Gamma ray, grb, ray bursts, bursts grbs, afterglow, swift, prompt emission, prompt, lorentz factor, fireball |
| sh12 | 964 | Black hole, horizon, hole solutions, quasinormal modes, rotating black, spacetime, five dimensional, charged black, hawking radiation, schwarzschild black |
| sh13 | 948 | Inflation, non gaussianity, slow roll, curvature perturbation, primordial, f nl, power spectrum, curvaton, inflationary models, bispectrum |
| sh2 | 3171 | Galaxies, star formation, redshift, formation rate, sample, early type, rest frame, active galactic, luminosity, stellar mass |

**Table 4** continued

| No. | Size | Word labels |
|---|---|---|
| sh3 | 2685 | Meson, decays, qcd, pi pi, quark, j psi, bar, pi, hadronic, lattice |
| sh4 | 2211 | Molecular cloud, protostellar, cores, massive star, toward, c 13, cloud, outflow, young stellar, h ii |
| sh5 | 1677 | Dark energy, universe, equation, cosmological, quintessence, type ia, phantom, matter, lambda cdm, scalar field |
| sh6 | 1454 | Standard model, higgs, minimal supersymmetric, supersymmetric standard, neutrino mass, lhc, lepton, seesaw, right handed, tev |
| sh7 | 1368 | Brown dwarf, pre main, tauri stars, dwarfs, main sequence, low mass, substellar, stars, young, spectral type |
| sh8 | 1283 | Asteroid, comet, main belt, kuiper belt, solar system, bodies, perihelion, orbits, nucleus, near earth |
| sh9 | 1145 | Eclipsing binary, star, orbital period, pulsation, wilson devinney, delta scuti, contact binary, mass transfer, light curves, photometric |
| sr0 | 158 | S106, bol2, z905, pphl7, loser, ecosystems, 3310, gurzadyan, ampicillin, eff1 |
| sr104 | 407 | Body problem, restricted three, periodic orbits, three body, equilibrium points, photogravitational, lyapunov, collinear, families, planar |
| sr1049 | 209 | Bacillus, subtilis, bacterial, spores, microorganisms, drilling, tinto, biological, rio, mars |
| sr106 | 3557 | Yang mills, gauge theory, mills theory, noncommutative, lattice, supergravity, string, qcd, finite temperature, branes |
| sr1084 | 112 | Elite, terraforming, lander, expeditions, lidov, philae, rendezvous, unsupported, simulant, society |
| sr126 | 19,988 | Galaxies, redshift, star formation, clusters, sample, active galactic, agn, sloan digital, digital sky, halo |
| sr138 | 109 | Nonextensive, tsallis, microcanonical, caloric, mechanics, inequivalence, additivity, bose, coarse grained, self gravitating |
| sr1381 | 477 | Meteor, leonid, geminid, radiant, nutation, ablation, video, trails, shower, perseid |
| sr147 | 260 | Rotational transitions, anion, ab initio, c6h, irc 10216, dissociative recombination, vibrational, tmc, molecule, franck |
| sr16 | 403 | Mars, deposits, hesperian, crater, amazonian, volcanic, hirise, geological, gullies, fluvial |
| sr17 | 33,874 | Star, main sequence, binary, light curve, gamma ray, white dwarf, neutron star, emission, low mass, x ray |
| sr191 | 3941 | Decays, meson, qcd, pi pi, j psi, leading order, bar, quark, inclusive, factorization |
| sr238 | 457 | Neutron capture, poor stars, nucleosynthesis, extremely metal, metal poor, capture elements, process elements, cemp, third dredge, isotopes |
| sr266 | 262 | Ozone, aerosol, aod, toms, stratospheric, envisat, retrieval, gome, sciamachy, modis |
| sr310 | 1733 | Ionospheric, gps, tec, iri, electron content, mesosphere, degrees n, total electron, ionosonde, summer |
| sr330 | 169 | Ethyl, rydberg, lih, formate, vibrational, ch3ch2cn, molecule, cyanide, predissociation, ab initio |
| sr355 | 2314 | Auroral, substorm, solar wind, magnetopause, magnetosphere, plasma sheet, field aligned, cluster spacecraft, ionospheric, ion |
| sr36 | 224 | Rainfall, tropical, precipitation, meteorological, mesoscale, mm5, ocean, sea level, grace, weather |
| sr365 | 102 | Sipm, astrosat, irst, fbk, ray astronomy, readout, counters, nct, mega, calorimeter |
| sr381 | 696 | Hanle, focal, mirrors, adaptive optics, wavefront, optics, integral field, laser guide, guide star, ifu |

**Table 4** continued

| No. | Size | Word labels |
|---|---|---|
| sr384 | 100 | Riemannian, osserman, anholonomic, lorentzian, finsler, causal, manifold, nilpotent, chronological, achronal |
| sr400 | 7076 | Solar, coronal, active region, cme, flare, sunspot, mass ejections, magnetic flux, quiet sun, transition region |
| sr403 | 4720 | Asteroid, saturn, comet, titan, cassini, jupiter, icarus, albedo, main belt, kuiper belt |
| sr425 | 2593 | Standard model, higgs, lhc, minimal supersymmetric, supersymmetric standard, lepton, seesaw, neutrino masses, right handed, leptogenesis |
| sr440 | 426 | Sail, thrust, debris, propulsion, restricted three, earth orbit, body problem, trajectory, maneuvers, geo |
| sr474 | 269 | Interferometric gravitational, wave detectors, mirrors, geo 600, gravitational wave, suspension, lcgt, interferometer, fused, thermoelastic |
| sr48 | 14,588 | Scalar field, spacetime, metric, inflation, cosmological constant, dark energy, gravity, general relativity, universe, einstein |
| sr5 | 136 | Plants, seedlings, arabidopsis, life support, wheat, grown, gravitropism, bioregenerative, shoots, germination |
| sr502 | 421 | r matrix, oscillator strengths, breit pauli, dielectronic recombination, transition probabilities, collision strengths, electron impact, 3p, impact excitation, rate coefficients |
| sr503 | 320 | Dynamos, dynamo action, magnetic reynolds, electromotive, reynolds number, prandtl number, geodynamo, magnetic helicity, nonhelical, magnetic prandtl |
| sr53 | 580 | Presolar, meteorites, isotopic compositions, chondrules, minerals, inclusions, isotopic, lunar, olivine, solar nebula |
| sr578 | 1424 | Pamela, matter annihilation, wimp, neutrino, weakly interacting, interacting massive, dark matter, positron, super kamiokande, theta 13 |
| sr628 | 131 | Tourism, economic, human, stiefel, industry, quaternions, kustaanheimo, social, countermeasures, psychological |
| sr644 | 516 | Pentaquark, nucleon, baryon, octet, decuplet, pion, n c, form factors, chiral, strangeness |
| sr792 | 126 | Dose, hzetrn, dosimetry, liulin, aircrew, shielding, tissue, space station, international space, station iss |
| sr972 | 844 | Ultra high, air shower, cosmic rays, extensive air, uhecr, pierre auger, 19 ev, auger observatory, energy neutrinos, neutrino flux |
| u1 | 18,259 | Galaxies, redshift, active galactic, agn, star formation, galactic nuclei, quasar, sample, gas, galaxy clusters |
| u10 | 4262 | Ray binary, x ray, neutron star, black hole, hard state, rossi x, timing explorer, ray timing, ultraluminous x, rxte |
| u11 | 3954 | Grb, ray bursts, gamma ray, afterglow, bursts grbs, sn, explosion, type ia, swift, supernova |
| u12 | 3522 | Pulsar, supernova remnant, psr, snr, neutron stars, wind nebula, anomalous x, radio pulsars, magnetar, remnant snr |
| u13 | 3096 | Ads, black holes, horizon, spacetimes, hole solutions, quasinormal modes, supergravity, dimensional, hawking radiation, anti |
| u14 | 2658 | Gravitational wave, lisa, inspiral, binary black, wave detectors, ligo, laser interferometer, numerical relativity, post newtonian, waveforms |
| u15 | 2171 | Planetary nebulae, pne, post agb, asymptotic giant, mira, central star, nebulae pne, agb stars, pn, symbiotic |
| u16 | 2087 | White dwarf, nova, cataclysmic variable, dwarf nova, subdwarf b, wd, sdb stars, orbital period, sdb, superhumps |
| u17 | 1272 | Auroral, substorm, magnetopause, ionospheric, cluster spacecraft, plasma sheet, magnetosheath, field aligned, superdarn, dayside |

**Table 4** continued

| No. | Size | Word labels |
|---|---|---|
| u18 | 861 | Ionospheric, iri, degrees n, tec, electron content, ionosonde, summer, total electron, fof2, winter |
| u19 | 233 | Body problem, restricted three, periodic orbits, three body, photogravitational, equilibrium points, sail, collinear, sitnikov, thrust |
| u2 | 12,432 | Inflation, dark energy, microwave background, cosmic microwave, cosmological, universe, scalar field, gravity, cmb, background cmb |
| u20 | 153 | Nutation, iau, celestial reference, p03, iers, cip, capitaine, celestial mechanics, chandler, mathews |
| u21 | 150 | Life support, plant, wheat, food, bioregenerative, crops, waste, biomass, cultivation, ecological |
| u3 | 11,477 | Star, planets, hd, main sequence, brown dwarfs, radial velocity, planet formation, transit, type stars, extrasolar planets |
| u4 | 7925 | Solar, coronal mass, active region, cme, flare, magnetic field, mass ejections, sunspot, quiet sun, chromosphere |
| u5 | 6324 | Mars, comet, asteroid, titan, saturn, cassini, albedo, icarus, jupiter, ice |
| u6 | 5692 | Standard model, neutrino, higgs, lhc, minimal supersymmetric, lepton, supersymmetric standard, gev, muon, top quark |
| u7 | 5277 | Molecular cloud, protostellar, cloud, interstellar, star forming, young stellar, molecules, forming region, massive star, c 13 |
| u8 | 5276 | Qcd, quark, meson, lattice, decays, chiral, pi pi, gluon, j psi, pion |
| u9 | 4685 | Globular clusters, fe h, metal poor, giant branch, red giant, metallicity, stars, milky way, dwarf spheroidal, galactic globular |

## Shared document sets

Table 5 lists for the 13 largest shared document sets and for each solution the cluster that incorporates the respective shared document set. We further provide the word based and thesaurus term based labels for each of the shared document sets in Table 6.

**Table 5** Shared document sets

| Set no. | # Documents | Associated clusters |
|---|---|---|
| 1 | 4162 | sr400 c3 u4 ok14 ol8 en11 eb4 hd8 |
| 2 | 3171 | sr126 c1 u1 ok12 ol11 en8 eb6 hd3 |
| 3 | 2685 | sr191 c7 u8 ok6 ol25 en2 eb13 hd10 |
| 4 | 2211 | sr17 c5 u7 ok24 ol2 en8 eb2 hd4 |
| 5 | 1677 | sr48 c2 u2 ok17 ol29 en5 eb11 hd5 |
| 6 | 1454 | sr425 c8 u6 ok5 ol20 en2 eb13 hd11 |
| 7 | 1368 | sr17 c4 u3 ok29 ol0 en1 eb2 hd4 |
| 8 | 1283 | sr403 c12 u5 ok22 ol12 en3 eb3 hd4 (hd7 hd9) |
| 9 | 1145 | sr17 c13 u3 ok2 ol6 en1 eb1 hd4 |
| 10 | 1083 | sr403 c14 u5 ok26 ol15 en3 eb3 hd4 (hd7 hd9) |
| 11 | 1066 | sr17 c15 u11 ok21 ol9 en4 eb8 hd6 |

**Table 5** continued

| Set no. | # Documents | Associated clusters |
|---|---|---|
| 12 | 964 | sr48 c10 u13 ok3 ol13 en9 eb10 hd5 |
| 13 | 948 | sr48 c2 u2 ok19 ol18 en5 eb11 hd5 |

**Table 6** Shared document sets indicating thematic cores

| No. | Word labels | Thesaurus term labels |
|---|---|---|
| Gravitation and Cosmology | | |
| 5 | Dark energy, universe, equation, cosmological, quintessence, type ia, phantom, matter, lambda cdm, scalar field | Dark energy, large scale structure of the universe, quintessence,cosmological models, type ia supernovae, observational cosmology, intergalactic medium, cosmic background radiation, scalar tensor vector gravity, general theory of relativity |
| 12 | Black hole, horizon, hole solutions, quasinormal modes, rotating black, spacetime, five dimensional, charged black, hawking radiation, schwarzschild black | Black holes, quasinormal modes, hawking radiation, event horizons, schwarzschild black holes, stars, black hole thermodynamics, relativity, naked singularities, gravitation |
| 13 | Inflation, non gaussianity, slow roll, curvature perturbation, primordial, f nl, power spectrum, curvaton, inflationary models, bispectrum | Non gaussianity, radio astronomy, cosmic microwave background radiation, spectral index, gravitational waves, large scale structure of the universe, cosmic strings, string theory, p branes, beyond the standard model |
| Atroparticle Physics | | |
| 3 | Meson, decays, qcd, pi pi, quark, j psi, bar, pi, hadronic, lattice | Light cones, perturbation methods, relativity |
| 6 | Standard model, higgs, minimal supersymmetric, supersymmetric standard, neutrino mass, lhc, lepton, seesaw, right handed, tev | Neutrino masses, beyond the standard model, supersymmetric standard model, supersymmetry, leptogenesis, grand unified theory, technicolor, origin of the universe, baryogenesis, solar neutrinos |
| Astrophysics (Galaxies) | | |
| 2 | Galaxies, star formation,redshift, formation rate, sample, early type, rest frame, active galactic, luminosity, stellar mass | Galaxies, redshift, doppler shift, star formation, milky way galaxy, stellar, evolution astronomical research, galaxy groups, active galactic nuclei, surveys |
| 11 | Gamma ray, grb, ray bursts, bursts grbs, afterglow, swift, prompt emission, prompt, lorentz factor, fireball | Gamma ray bursts, stellar phenomena, fireballs, astroparticle physics, light curves, neutron stars, ejecta, photometry, magnetars, collimation |
| Astrophysics (Stars) | | |
| 4 | Molecular cloud, protostellar, cores, massive star, toward, c 13, cloud, outflow, young stellar, h ii | Molecular clouds, interstellar medium, gaseous spheres, protostars, nebulae, clouds, young stellar objects, h ii regions, infrared astronomical satellite, planetary atmospheres |
| 7 | Brown dwarf, pre main, tauri stars, dwarfs, main sequence, low mass, substellar, stars, young, spectral type | Brown dwarfs, t tauri stars, circumstellar matter, pre main sequence stars, low mass stars, dwarf stars, stellar classification, stars, classical t tauri stars, proper motions |

**Table 6** continued

| No. | Word labels | Thesaurus term labels |
|---|---|---|
| 9 | Eclipsing binary, star, orbital period, pulsation, wilson devinney, delta scuti, contact binary, mass transfer, light curves, photometric | Eclipsing binary stars, peculiar objects, binary systems, ap stars, chemically peculiar stars, radial velocity, mass transfer, light curves, main sequence stars, observation techniques |
| Solar Physics | | |
| 1 | Solar, magnetic field, coronal, flare, plasma, active region, cme, mass ejections, reconnection, euv | Solar physics, stellar structure, magnetic fields, coronal mass ejections, coronae, solar flares, solar corona, sunspots, starspots, chromosphere |
| Planetary Science | | |
| 8 | Asteroid, comet, main belt, kuiper belt, solar system, bodies, perihelion, orbits, nucleus, near earth | Asteroids, comets, solar system, near earth objects, orbits, trans neptunian objects, kuiper belt, perihelion, comas, centaurs |
| 10 | Titan, cassini, atmosphere, mars, saturn, methane, haze, aerosol, ice, surface | Natural satellites, saturnian satellites, solar system, planets, mars, galilean satellites, planetary atmospheres, saturn, methane, stratosphere |

## Details of analysis of local data versus global data

Solution *sr* represents the topical structure of some domains as highly concentrated with the large majority of documents included in a small number of topic clusters: in *Astrophysics* 98% of documents are contained in one large cluster, sr48; in *Gravitational Physics, Cosmology* the two largest clusters cover 96% of documents, and in *Solar Physics* all documents are included in a single cluster.[13] In the other three domains documents are less concentrated and spread across a larger set of topics: in *Astroparticle Physics* the 2 largest clusters account for only 58% of documents in the domain, to reach a coverage of 96% the 5 largest clusters need to be combined; in *Planetary Science* the largest cluster covers 74% of documents in the domain, and only the four largest clusters account for 97% of the documents; in *Space Science* the 2 largest clusters cover 64% of the documents in the domain and only the 10 largest clusters combined account for at least 96% of the documents.

The variation in the resolution of topics by domain is visible also in the lexical fingerprint analysis: At one end of the extreme, in *Gravitational Physics, Cosmology* (Fig. 11), the *sr* solution looks most similar to another low resolution solution, namely *eb* that distinguishes merely two topics: one on cosmology and inflation, and one on gravitational waves; *eb* slightly differs from *sr* in that it subsumes also black holes and spacetime into this latter topic. In the domain of *Solar physics* (Fig. 12) and its extension to

[13] The size of a domain used in this calculation is determined after assigning the 35 largest clusters in the *sr* solution to domains. They represent 97.5% of all documents covered by the *sr* solution and are described in detail in a Table in Boyack (2017a).

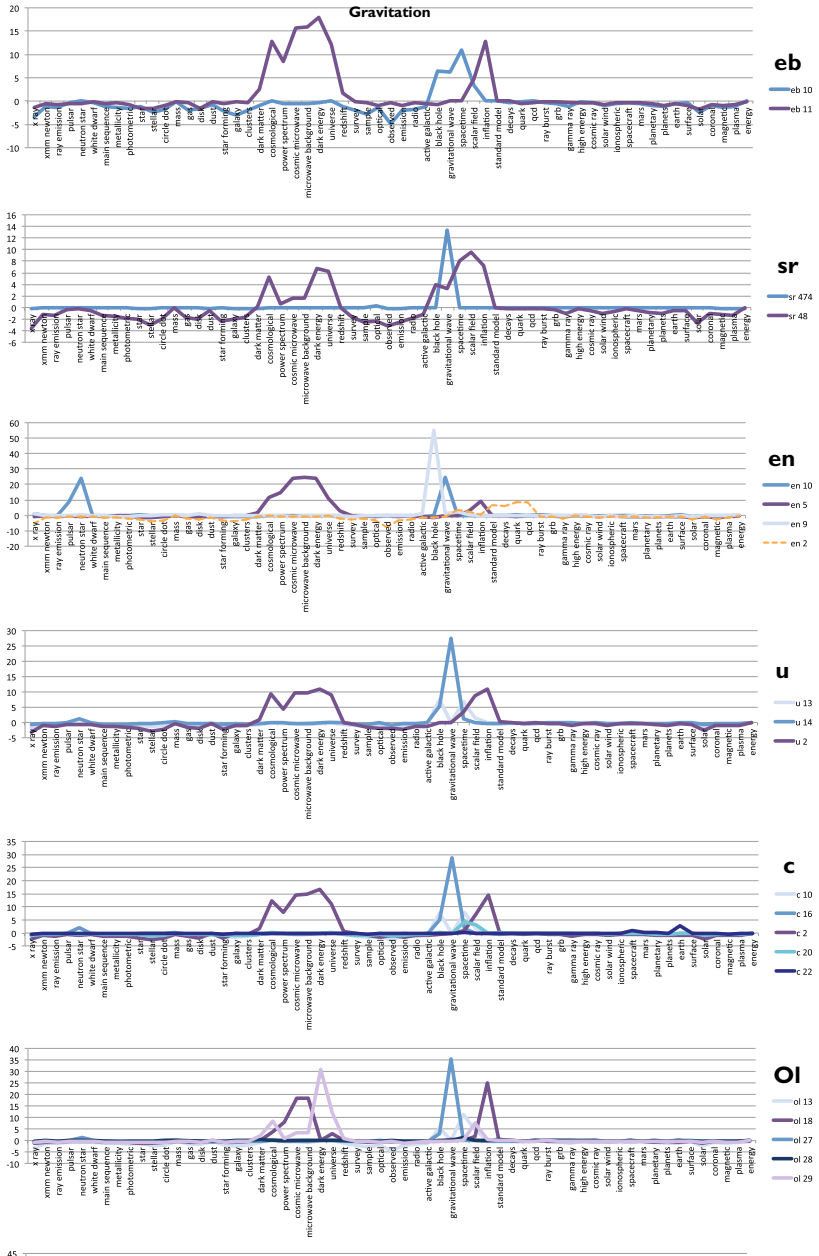



**Fig. 11** Lexical fingerprint: Gravitational Physics and Cosmology

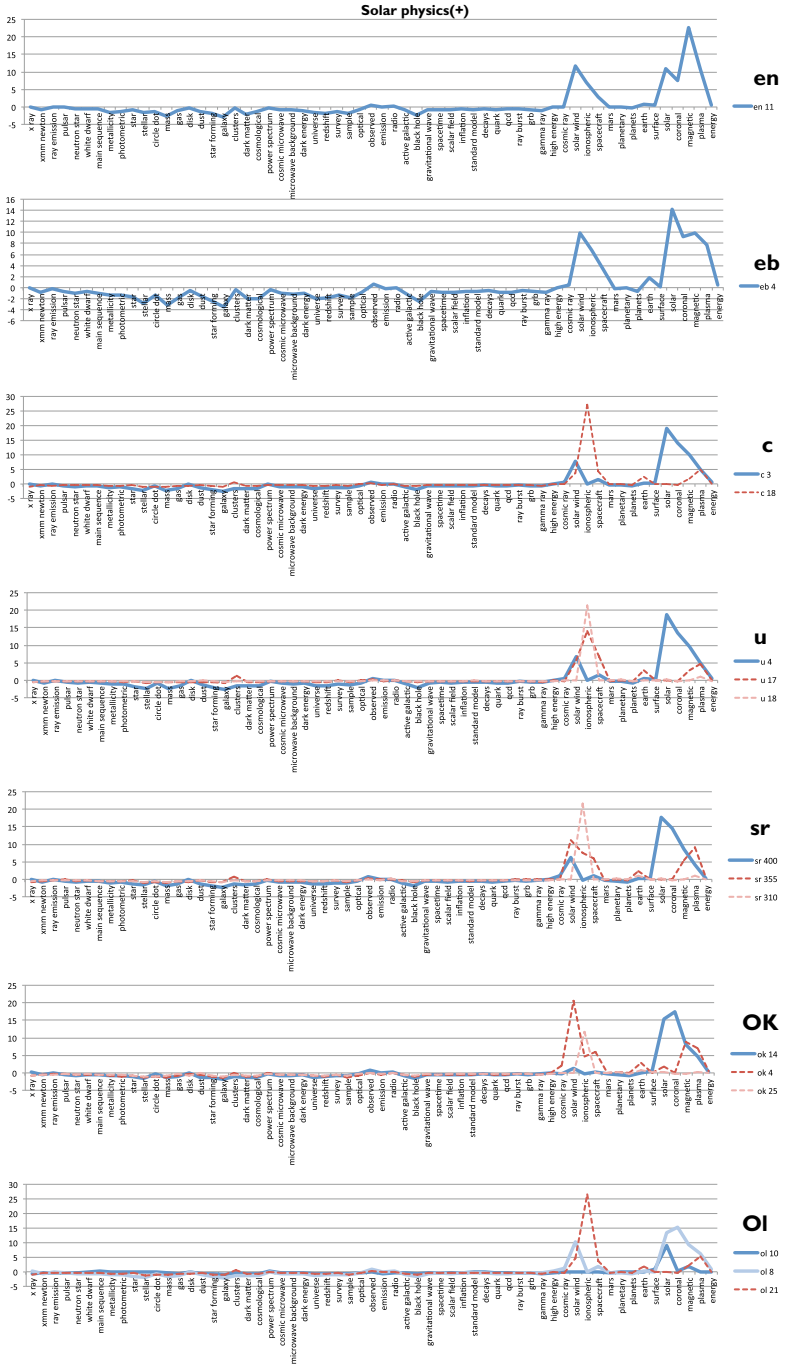



**Fig. 12** Lexical fingerprint: Solar physics. Additional (*dotted*) lines shown represent topics that have been assigned to the domain of *Space Science* based on journal signature analysis. They relate to effects of the sun on the earth (aurora, ionosphere)

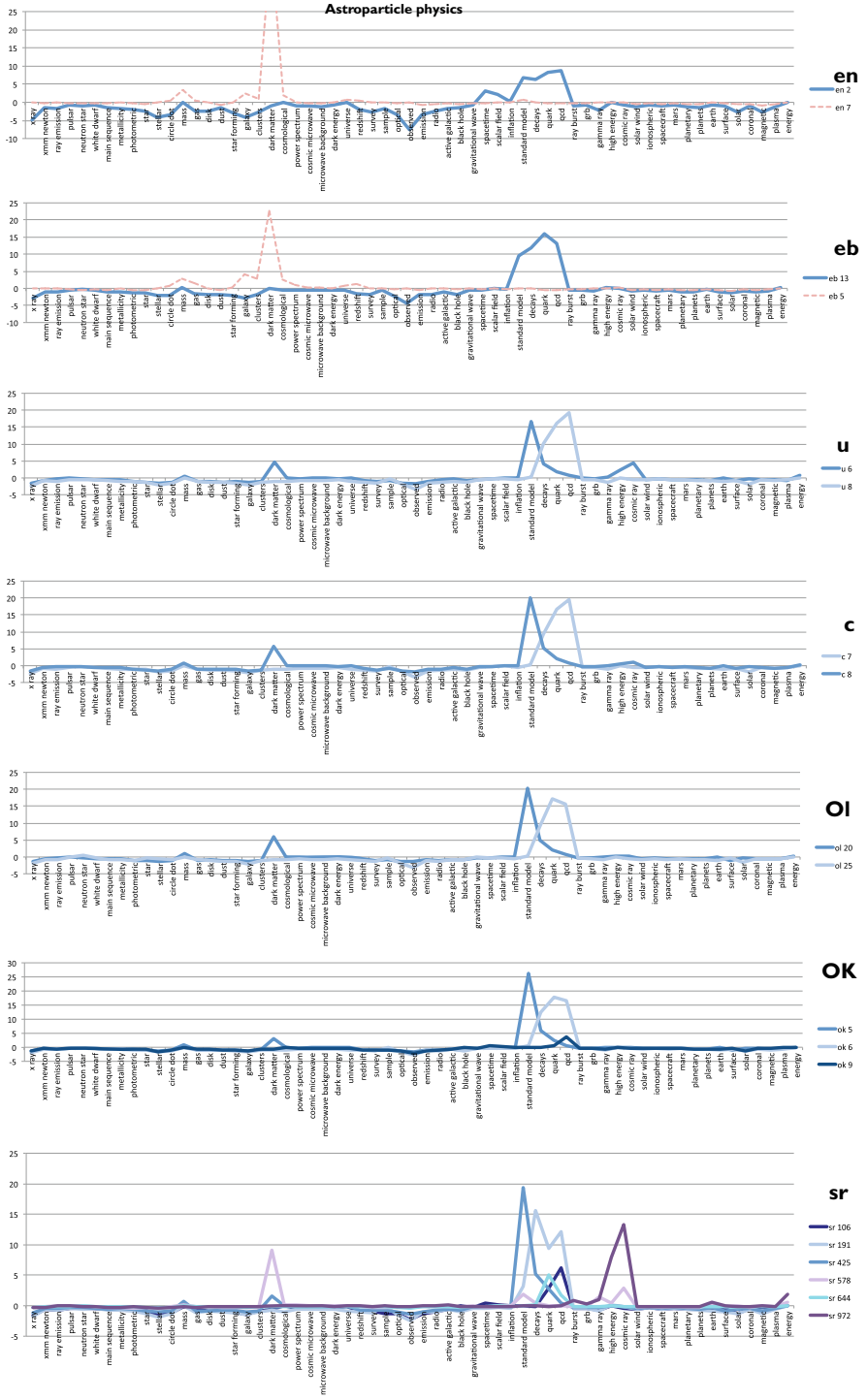



**Fig. 13** Lexical fingerprint: Astroparticle physics

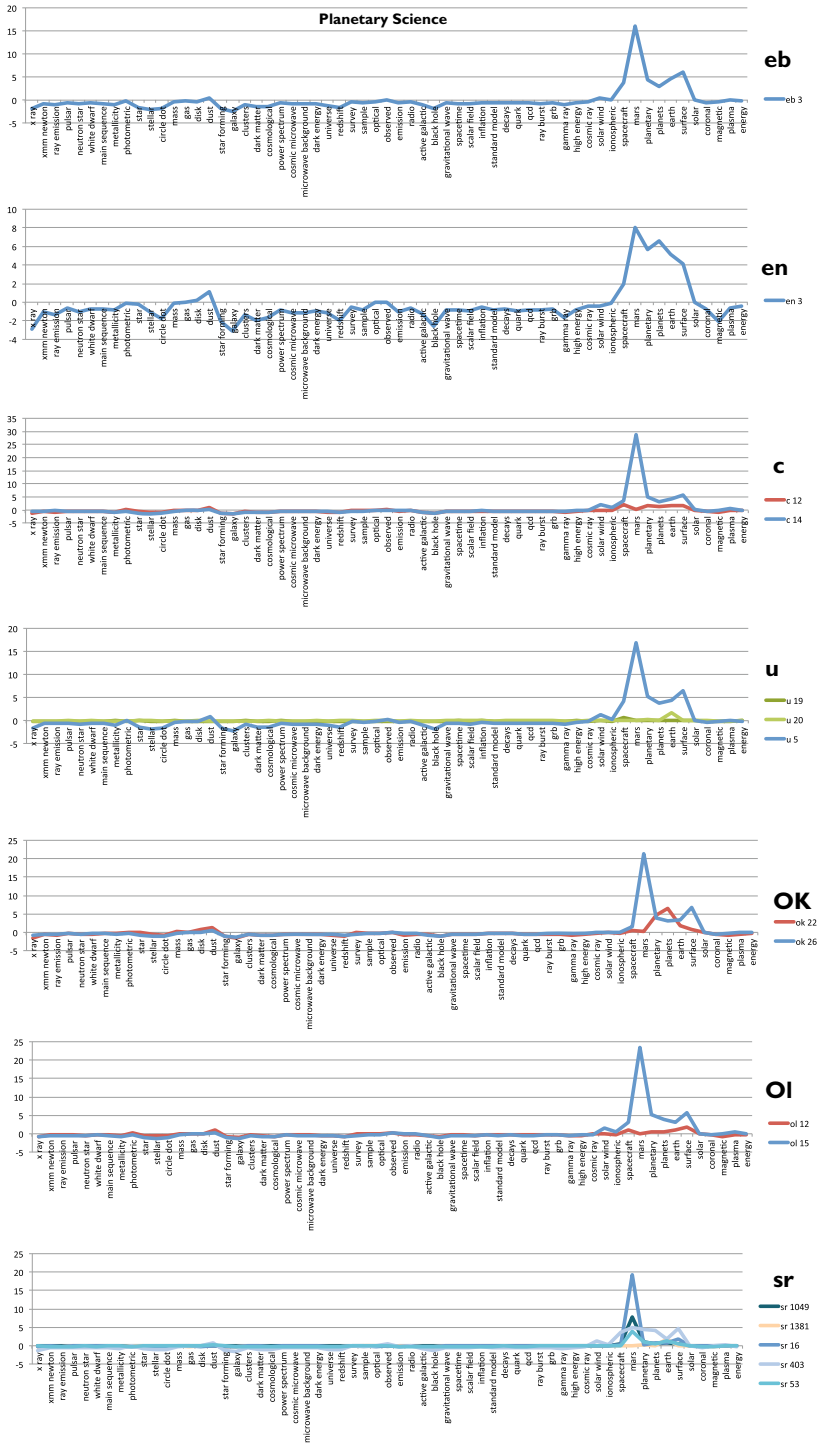



**Fig. 14** Lexical fingerprint: Planetary science

sun related topics in *Space Science* the topics identified by *sr* are very similar to the fingerprint of the topics identified by *c*, *u*, and *ok*. Finally, at the other end of the extreme, in *Astro Particle Physics* and *Planetary Science* (Figs. 13, 14), *sr* adds considerable resolution to the topics identified compared to the other six solutions.

## Details of analysis of citation based versus semantic data models

*Direct citation* (*c*, *u*) *versus hypersemantic data model* (*ol*, *ok*)

The labels of the two topics that solutions *ol* and *ok* detect and that are largely missing from the citation based solutions *c* and *u*, are:

- ok27: "performance, scientific, technology, mission, astronomical, development, research, flight, cost, software",
- ol16: "performance, human, research, scientific, development, technology, earth, mission, astronomical, control".

We interpret these labels as suggesting a topic relating to space missions.

*Bibliographic coupling* (*eb*) *versus hybrid* (*en*)

A detailed analysis of the topical structure as shown by the affinity networks of the two solutions, see Fig. 9 reveals some distinct differences:

| | |
|---|---|
| Relative topic sizes | The three single topics that represent an entire domain in each solution seem blown up in *en* versus *eb*. The Astroparticle Physics topic en2 with 16.1% is almost twice in relative size than the corresponding topic eb13 (8.8%); same for the Planetary Science topics with a relative size of en3 of 12.4% versus a relative size of eb3 of 7.0%. The Solar Physics topic in *en* is about 15% larger (en11, 13.6%) versus (eb4, 11.7%) in *eb*. |
| Granularity by domain | *en* depicts the *Gravitational Physics, Cosmology* domain with greater granularity than *eb* does with only two large clusters for this domain. Both solutions agree in identifying a 'cosmology' topic, however they differ in the other topics. Solution *en* distinguishes further 'black holes' and 'gravitational wave sources' whereas *eb* identifies a comprehensive topic 'spacetime' that incorporates supergravity, as well as black holes and gravitational waves. In turn, *eb* depicts the *Astrophysics* domain with greater granularity than *en*. Both solutions identify five topically similar clusters on 'galaxy' (en8, eb6),'stars' (en1, eb1), 'dark matter' (en7, eb5), 'x-ray' (en6, eb9) and 'gamma ray' (en4, eb8). In addition, *eb* distinguishes clusters relating to 'accretion disks' (eb12), 'infra-red sources' (eb2) and 'star clusters' (eb7), see Fig. 15. |

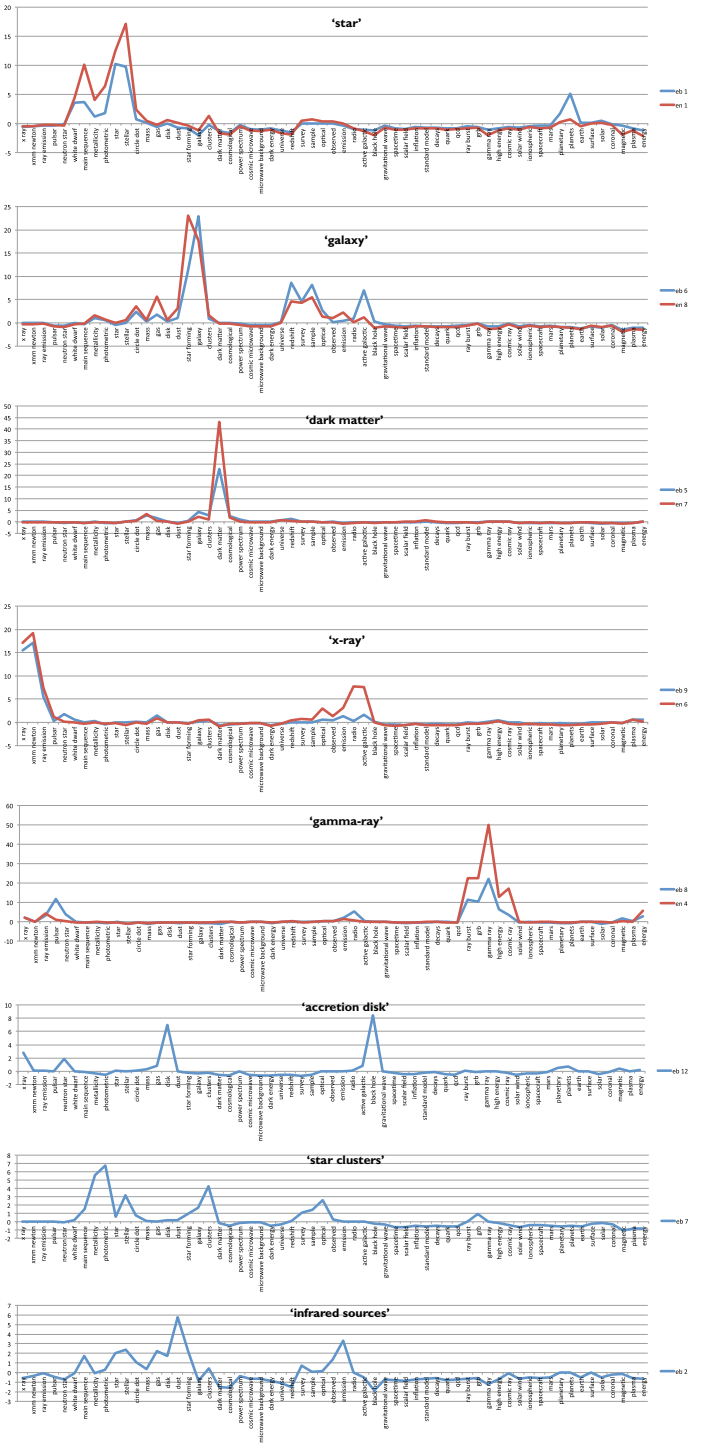



**Fig. 15** Comparison of fingerprints for *Astrophysics* topics in solutions *en* and *eb*

| | |
|---|---|
| Connectivity | The affinity network for solution *en* shows the domain of *Gravitational Physics and Cosmology* connecting to *Astrophysics* through two distinct bridges: on the one hand a large-scale structure, or cosmological theme, that spans 'cosmology' (en5), 'dark matter' (en7), and 'galaxies' (en8), and on the other hand through compact objects such as 'black holes ' (en9) and other 'sources of gravitational waves' (en10). Solution *eb* replicates the large-scale structure, or cosmological theme as a main connection between *Gravitational Physics and Cosmology* and *Astrophysics*. However it does not expose the second, compact objects bridge as a distinct feature: the topic of black holes is subsumed in the large 'spacetime' topic cluster eb10, and in contrast to *en* the interlinking to the only other *Gravitational Physics and Cosmology* topic (eb11) is stronger than the strongest link over into the *Astrophysics* domain, to eb12 'accretion disks'. A striking feature of solution *en* is the way that the *Solar Physics* topic is connected, with its strongest links into 'gamma-ray', 'x-ray', and 'gravitational wave' sources. The two strongest links of the *Solar Physics* topic in *eb* link to 'stars' (eb1) and 'x-ray' (eb9). |

A number of the distinctive features of solution *en* relative to *eb* could be due to aggregation effects due to the lexical component in the data model: First, the extreme sizes of the *Planetary Science* topic and the *Astroparticle Physics* topic: A keyword occurrence search for 'exoplanets' in titles shows for *eb* a high concentration of this term in the topic eb1 'stars' in the *Astrophysics* domain. By contrast, for *en* the occurrence of this terms is split between two topics in *Planetary Science* and *Astrophysics*: en1 'stars' and en3 'solar system'—so a larger proportion of it has been aggregated into the single topic en3 in *Planetary Science*. This is further corroborated by the broader and stronger signal for 'planetary' and 'planets' in the lexical fingerprint of en3 (see Fig. 14). From a subject expert's perspective, the search for extra-solar planets is to a large extent about the close observation of stars and variations in their movements or in their radiation. Hence we can expect publications on the search for extra solar planets to tie in tightly with the literature on stellar observations. However, this connection may be weakened when lexical terms relating to 'planets' are taken into account such that links into the planetary science literature gain greater weight leading to a partially stronger integration of publications on extra-solar planets with the topic of 'planets'. We speculate that a similar effect may be at work when aggregating topics into the *Astroparticle Physics* topic. Based on the entropy based labeling with thesaurus terms, solution *en* integrates the theme of supergravity into *Astroparticle Physics* (en2) whereas *eb* integrates it into the 'spacetime' topic (eb10). This corresponds to en2 having a peak in its lexical fingerprint for 'spacetime' (and no such peak in any of the three topics in *Gravitational Physics and Cosmology* (see Figs. 11, 13), whereas *eb* has a peak for the term spacetime only in its lexical fingerprint for the 'spacetime' topic (eb10) in *Gravitational Physics and Cosmology*. It is not clear what lexical terms are responsible, but there is the possibility that the literature of supergravity is using more frequently terms that are used also in other field theories that are part of the *Astroparticle Physics* domain, and hence the lexical approach emphasizes the link from supergravity to other themes in *Astroparticle Physics*.

Further, we find that the term plasma is relatively concentrated in *en* with 71% of occurrences in the single *Solar Physics* topic (en11). In *eb* the concentration of this term in the *Solar Physics* topic (eb4) is lower, at 53%. Compared to eb4 in solution *eb*, en11 in

solution *en* has relatively stronger links to the topics of specific types of radiation sources, namely 'gamma-ray sources' (en4), 'x-ray sources' (en6), and 'gravitational wave sources' (en10). This suggests that due to the lexical component in the data model, documents of these respective topics that use terms like 'plasma' have been merged with the solar physics topic.

## Details of analysis of local versus global clustering

As explained in the main text we explore the differences between the topics found by the two solutions by comparing the lexical fingerprints of topics in *hd* versus *c*, selecting as examples topics in the two domains of *Astroparticle Physics* (see Fig. 10) and *Gravitational Physics and Cosmology* (see Fig. 16). Note that since no journal-signature based assignment of topic clusters to domains is available for solution *hd*, the selection of topics for these two domain-wise comparisons proceeded as follows: we identified those terms that scored high in the domain-wide fingerprint analysis of the seven global clustering

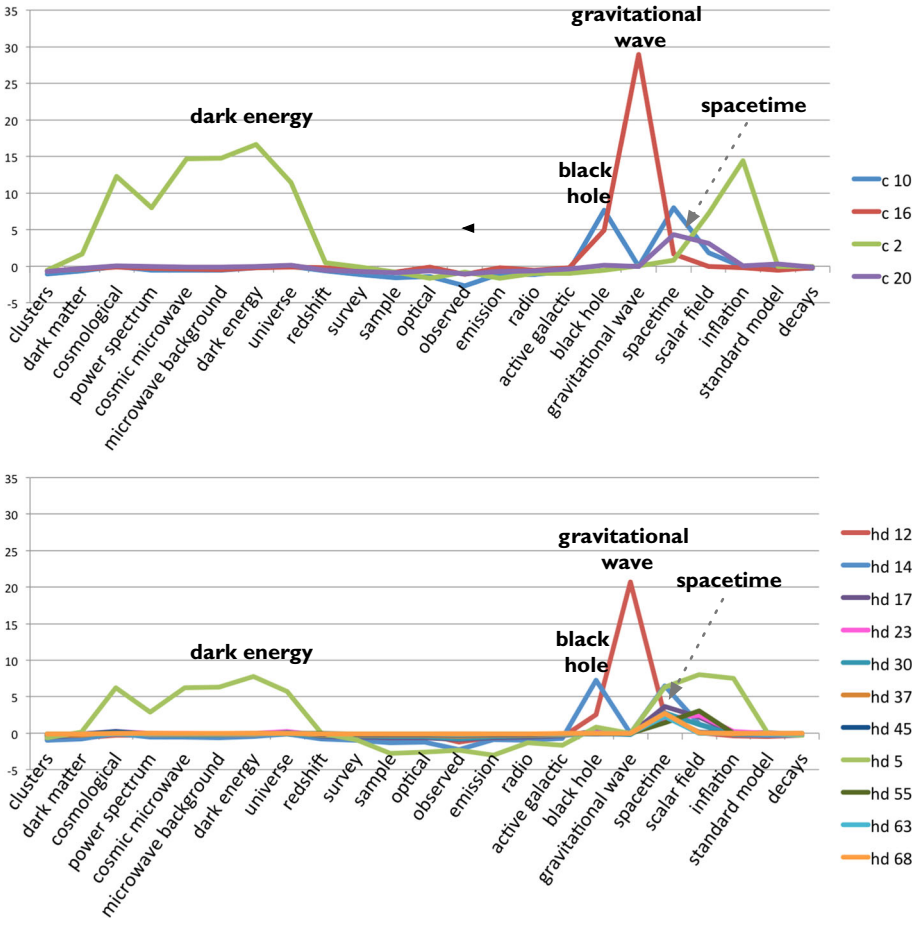



**Fig. 16** Comparison of (partial) fingerprints for *Gravitational Physics and Cosmology* topics in solutions *hd* and *c*

solutions depicted in Fig. 13, and Fig. 11. We then determined those topics in solutions *hd* and *c* that had a major peak for those terms. For example, although cluster hd6 has a positive score for the term 'black hole', it was not considered as part of the domain *Gravitational Physics and Cosmology* because this score is only a minor peak compared to the much higher scores for terms such as 'gamma-ray', 'x-ray', and 'neutron star' that relate more strongly to the *Astrophysics domain* (see data file clusterfactors.csv, doi: http://dx.doi.org/10.17026/dans-zzq-z4xh).

When comparing fingerprints of topics in *Astroparticle Physics* for solutions *hd* and *c* we observe agreement between the solutions regarding two major topics, and one extra topic that only *hd* detects. The other topics identified by *hd* seem to be, based on their lexical fingerprint, variations of these three topics and tend to be smaller: hd20 (1920 documents) and hd27 (1296 documents) with a double peak at 'standard model' and 'qcd', hd21 (1512 documents) with a broader peak at 'decays' and 'quark', hd26 (1133 documents) and hd35 (499 documents) with their highest peak at 'quark', and finally hd19 (1342 documents), hd40 (478 documents), hd29 (1051 documents), hd42 (474 documents), hd43 (269 documents), hd49 (325 documents), hd56 (246 documents) with their major peak at 'qcd'.

Comparing topics detected by solutions *hd* and *c* in the domain *Gravitational Physics and Cosmology*, we we find that the two solutions agree in the identification of four major topics that have major peaks at 'cosmological/cosmic microwave/dark energy/inflation' (*c2*: 8954 documents, *hd5*: 21873 documents), 'gravitational wave' (*c16*: 3156 documents, *hd12*: 4193 documents), 'black hole/spacetime' (*c10*: 3904 documents, *hd14*: 3887 documents), and 'spacetime/scalar field' (*c20*: 1963 documents, *hd17*: 1793 documents). The additional topics that hd identifies in the domain *Gravitational Physics and Cosmology* all seem to be variants of hd 17 of smaller size (ranging between 1148 and 188 documents), with hd37, hd63 and hd 68 having a positive score only for 'spacetime' and hd23, hd30, hd45, hd55 having positive scores for 'spacetime' and 'scalar field'.

Striking is the much bigger size of hd5 relative to c2, with hd5 including twice as many documents than c2. Seemingly, the fact that *hd* allows for multiple assignments of documents to several topics led to a large number of documents that are assigned to other topics in *c* to be included in hd5. A direct of comparison of their full lexical fingerprints (see Fig. 17) suggests that hd5 captures additional themes relating to the terms 'spacetime', 'black hole', and 'ionospheric'.

The entropy based labels of the two document clusters compare as follows (distinct terms highlighted in bold):

- hd 5 (words) "scalar field", "dark energy", "inflation", "**cosmological constant**", "**gravity**", "**spacetime**", "microwave background", "cosmic microwave", "brane", "universe"
- hd5 (thesaurus terms) "beyond the standard model", "cosmic microwave background radiation", "p branes", "dark energy", "**relativity**", "cosmological models", "radio astronomy", "**quantum gravity**", "**general theory of relativity**", "**gravitational singularities**"
- c2 (words)"dark energy", "microwave background", "cosmic microwave", "inflation", "**cosmological**", "universe", "cmb", "power spectrum", "background cmb", "scalar field"
- c2 (thesaurus terms)"cosmic microwave background radiation", "radio astronomy", "dark energy", "**large scale structure of the universe**", "cosmological models",

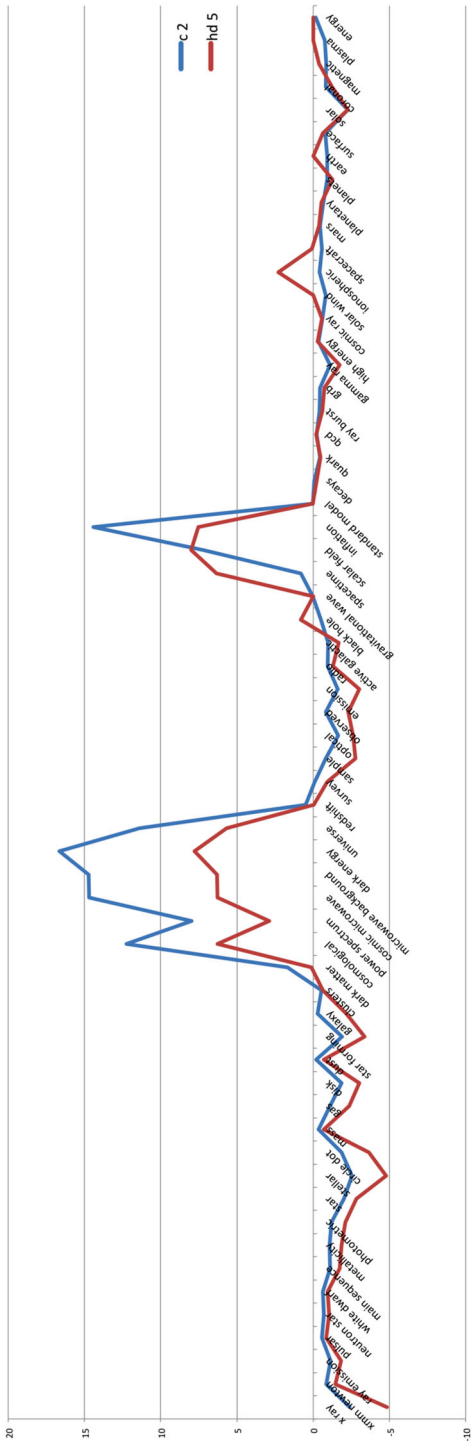



**Fig. 17** Comparison of full fingerprints for topics hd5 and c2

"**non gaussianity**", "p branes", "beyond the standard model", "**quintessence**", "**observational cosmology**"

These differences are not easy to interpret for the non-expert but may suggest that c2 is more focused on observational cosmology, whereas hd5 includes a larger portion of theoretical literature on quantum gravity.

## References


Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), 10,008.

Boyack, K. (2017a). Investigating the effect of global data on topic detection. In J. Gläser, A. Scharnhorst & W. Glänzel (Eds.), *Same data—Different results? Towards a comparative approach to the identification of thematic structures in science*. Special Issue of Scientometrics. doi:10.1007/s11192-017-2297-y.

Boyack, K. W. (2017b). Thesaurus-based methods for mapping contents of publication sets. In J. Gläser, A. Scharnhorst & W. Glänzel (Eds.), *Same data—Different results? Towards a comparative approach to the identification of thematic structures in science*. Special Issue of Scientometrics. doi:10.1007/s11192-017-2304-3.

Boyack, K. W., Glänzel, W., Gläser, J., Havemann, F., Thijs, B., Van Eck, N.J., et al. (2017). Topic identification challenge. In J. Gläser, A. Scharnhorst & W. Glänzel (Eds.), *Same data: Different results? Towards a comparative approach to the identification of thematic structures in science*. Special Issue of Scientometrics. doi:10.1007/s11192-017-2307-0.

Burger, M., & Bujdosó, E. (1985). Oscillating chemical reactions as an example of the development of a subfield of science. In J. R. Field & M. Burger (Eds.), *Oscillating and traveling waves in chemical systems* (pp. 565–604). New York: Wiley.

Dillo, I., van Horik, R., & Scharnhorst, A. (2013). Training in data curation as service in a federated data infrastructure-the frontoffice–backoffice model. In: *International conference on theory and practice of digital libraries* (pp. 205–215). Springer.

Glänzel, W., & Thijs, B. (2017). Using hybrid methods and 'core documents' for the representation of clusters and topics. The astronomy dataset. In J. Gläser, A. Scharnhorst & W. Glänzel (Eds.), *Same data—Different results? Towards a comparative approach to the identification of thematic structures in science*. Special Issue of Scientometrics. doi:10.1007/s11192-017-2301-6.

Gläser, J. (2006). *Wissenschaftliche Produktionsgemeinschaften: Die soziale Ordnung der Forschung* (Vol. 906). Frankfurt am Main: Campus.

Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data: Different results? Towards a comparative approach to the identification of thematic structures in science. In J. Gläser, A. Scharnhorst & W. Glänzel (Eds.), *Same data—Different results? Towards a comparative approach to the identification of thematic structures in science*. Special Issue of Scientometrics. doi:10.1007/s11192-017-2296-z.

Havemann, F., Gläser, J., & Heinz, M. (2017). Memetic search for overlapping topics based on a local evaluation of link communities. In J. Gläser, A. Scharnhorst & W. Glänzel (Eds.), *Same data—Different results?Towards a comparative approach to the identification of thematic structures in science*. Special Issue of Scientometrics. doi:10.1007/s11192-017-2302-5.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., & Rafols, I. (2015). Bibliometrics: The leiden manifesto for research metrics. *Nature*, *520*, 429–431.

Koopman, R., & Wang, S. (2017a). Clustering articles based on semantic similarity. In J. Gläser, A. Scharnhorst & W. Glänzel (Eds.), *Same data—Different results? Towards a comparative approach to the identification of thematic structures in science*. Special Issue of Scientometrics. doi:10.1007/s11192-017-2298-x.

Koopman, R., & Wang, S. (2017b). Mutual information based labelling and comparing clusters. In J. Gläser, A. Scharnhorst & W. Glänzel (Eds.), *Same data—Different results? Towards a comparative approach to the identification of thematic structures in science*. Special Issue of Scientometrics. doi:10.1007/s11192-017-2305-2.

Koopman, R., Wang, S., & Scharnhorst, A. (2017). Contextualization of topics: Browsing through the universe of bibliographic information. In J. Gläser, A. Scharnhorst & W. Glänzel (Eds.), *Same data—Different results? Towards a comparative approach to the identification of thematic structures in science*. Special Issue of Scientometrics. doi:10.1007/s11192-017-2303-4.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge: Cambridge University Press.

Mayr, P., & Scharnhorst, A. (2015). Scientometrics and information retrieval: Weak-links revitalized. *Scientometrics*, *102*(3), 2193–2199.

Petersen, A. C. (2012). *Simulating nature: A philosophical study of computer-simulation uncertainties and their role in climate science and policy advice*. Boca Raton: CRC Press.

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, *105*(4), 1118–1123.

Van Eck, N. J., & Waltman, L. (2017). Citation-based clustering of publications using CitNetExplorer and VOSviewer. In J. Gläser, A. Scharnhorst & W. Glänzel (Eds.), *Same data—Different results?Towards a comparative approach to the identification of thematic structures in science*. Special Issue of Scientometrics. doi:10.1007/s11192-017-2300-7.

Velden, T., & Lagoze, C. (2013). The extraction of community structures from publication networks to support ethnographic observations of field differences in scientific communication. *Journal of the American Society for Information Science and Technology*, *64*(12), 2405–2427.

Velden, T., Yan, S., & Lagoze, C. (2017). Mapping the cognitive structure of astrophysics by infomap clustering of the citation network and topic affinity analysis. In J. Gläser, A. Scharnhorst & W. Glänzel (Eds.), *Same data—Different results?Towards a comparative approach to the identification of thematic structures in science*. Special Issue of Scientometrics. doi:10.1007/s11192-017-2299-9.

Waltman, L., & Eck, N. J. (2012). 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.

Waltman, L., & van Eck, N. J. (2013). A smart local moving algorithm for large-scale modularity-based community detection. *The European Physical Journal B*, *86*(11), 1–14.

Xie, P., & Xing, E. P. (2013). Integrating document clustering and topic modeling. arXiv preprint arXiv:13096874.