Validation of the Astro dataset clustering solutions with external data

Paul Donner*

This is a pre-print of an article published in Scientometrics. The final authenticated version is available online at: https://doi.org/10.1007/s11192-020-03780-3.

Abstract

We conduct an independent cluster validation study on published clustering solutions of a research testbed corpus, the *Astro* dataset of publication records from astronomy and astrophysics. We extend the dataset by collecting external validation data serving as proxies for the latent structure of the corpus. Specifically, we collect (1) grant funding information related to the publications, (2) data on topical special issues, (3) on specific journals' internal topic classifications and (4) usage data from the main online bibliographic database of the discipline. The latter three types of data are newly introduced for the purpose of clustering validation and the rationale for using them for this task is set out. We find that one solution based on the global citation network achieves better results than the competitors across three validation data sources but that another solution based on bibliographic coupling performs best on the special issues data.

Keywords

cluster validation; document clustering; structural bibliometrics

1 Introduction

Given the huge scale of published research and the accelerating growth of new publications, automatic grouping of documents, that is, clustering, has emerged as an important task in structural bibliometrics. Clustering scientific document collections has applications in information retrieval, knowledge organization and field delineation, field normalization of indicators, and the visualization or mapping of document collections. Scientific document clustering is by now a well established subdiscipline of bibliometrics. New methods are developed and older methods are improved constantly. Until recently, there was no standard dataset on which such contributions could be benchmarked against prior state-of-the-art results. However, with the introduction of the *Astro* dataset, such a central testbed is now available to the community (Gläser, Glänzel, & Scharnhorst, 2017). We contribute to this development by complementing the publication records dataset with external validation data. We also study the performance of the publicly released clustering solutions of the *Astro* dataset[1] with these validation datasets to gain important insights into the strengths and weaknesses

[*] Paul Donner, ORCID 0000-0001-5737-8483, Deutsches Zentrum für Hochschul- und Wissenschaftsforschung, Schützenstraße 6a, 10117 Berlin, Germany

[1] see http://www.topic-challenge.info

of the underlying document clustering approaches. Thus the major contribution of this study is to show that it is possible to judge the quality of publication clustering solutions, as operationalized by the correspondence of solutions to latent topic structure reflected in several different validation datasets.

The introduction proceeds with a discussion of clustering. We outline the motivation for this study and its significance and discuss the background of the clustering comparison initiative of which we evaluate the clustering solutions, followed by a discussion of the literature on evaluation of bibliometric clustering. In the second section we summarize the studied clustering solutions, present the collected external validation datasets, and justify their suitability for comparison of the performance of the clustering solutions. In method section we explain our choice of evaluation criterion measure. Next we present the results of the study. We conclude with a discussion of the findings.

1.1 Clustering

Data clustering is a statistical approach to organizing objects into coherent groups, called clusters, without any training data. The grouping structure is obtained from the similarities or explicit connections among the input objects. The task is to create groups such that within a group the objects are similar while the objects of different groups are not similar. Within the discipline of bibliometrics, the clustering of documents (by means of their metadata records and sometimes fulltexts) on the basis of their textual and relationship information (i.e. citations, common authors) is often needed to obtain a structural view of document sets without manual work. Other entities besides publications which are often clustered in bibliometrics are authors and journals. Publication-level clusterings provide an automatic organization of literature that does not require any a priori classifications, based on the information contained in document records or the documents themselves and their relations to other documents. Moreover, it allows for the reproducible construction of research field structural descriptions by specified criteria.

1.2 Motivation and significance

The inherent intellectual structure of science and the consequences of this structure are an ongoing and pervasive concern in bibliometrics. The basic fact that research necessarily builds on deliberately selected prior research creates the structures that are perceived by people and taken for granted when using such concepts as scientific fields, disciplines, specialties, or topics. As it is crucial to take these structures into account in bibliometric studies in order to obtain valid and robust results, a specialized subfield has developed which Gläser et al. (2017) refer to as *structural bibliometrics*. Publication and citation practices vary widely between structural units which restricts comparative studies unless the structure is explicitly taken into account. To this end, structural bibliometrics is researching methods of structure recovery such as clustering of scientometric units, science mapping and field delineation. An important application is the field-normalization of indicators in studies covering more than a single, well-defined structural area. Here, an ongoing line of research has recently established the superiority of article level clustering to conventional journal classification systems as organizing representations of the research system (Klavans & Boyack, 2017; Ruiz-Castillo & Waltman, 2015; Shu et al., 2019). Publication-level clustering can also serve as an input to science mapping and can be used as the basis for field delineation, which indicates this method's central position within structural bibliometrics.

The structures obtained in publication clustering need to be validated in order to judge the utility of the applied method. The present contribution aims to contribute to the validation of clustering approaches by applying new external validation data to published clustering solutions.

1.3 Topic Extraction Challenge

The Topic Extraction Challenge was a collaboration of scientometrics research groups active in clustering of publications into topics, that is, clusters of thematically closely related publications. As various approaches for this task have been developed, this collaboration sought to compare these approaches by their solutions for the same benchmarking dataset. A common set of publications was created for the field of astronomy and astrophysics based on original data from the Web of Science database, the *Astro* dataset. Methods and results of the involved research groups were published in a special issue of Scientometrics (volume 111, issue 2, May 2017). The specific purpose of the project was to learn to what extent the topics are independent of the particular clustering approaches and therefore are unlikely to be mere methodological artifacts. The Topic Extraction Challenge, besides its immediate results (Velden, Boyack, et al., 2017) has led to the unique opportunity that there are now publicly available a number of alternative solutions to a single original dataset which can be further explored. This we take as the point of departure for the study presented here in which the results of the different approaches are further compared using new external validation data.

1.4 Literature review

In this section we discuss the more recent literature on clustering of publications for bibliometric studies, with special consideration for the types of clustering validation. Most studies have used internal and external validation strategies. Internal validation refers to the use of statistical indices for characterizing how well the clustering property is fulfilled, that is, high within-cluster homogeneity and between-cluster separation. The advantage of internal validation is that no ground truth data is required (Halkidi, Vazirgiannis, & Hennig, 2015). Šubelj, van Eck, & Waltman (2016) compare various clustering algorithms on the same datasets with a number of such statistics. The major disadvantage is that clustering algorithms use different operationalizations of the clustering property, some of which are also used as evaluation statistics, which leads to biased results. Šubelj et al. (2016) encounter this problem. The authors point out that when they use modularity Q as an evaluation criterion, the Louvain algorithm excels because it directly optimizes Q. Likewise, the Infomap algorithm obtains the best value for the log-likelihood evaluation statistic because its method is based on optimizing a likelihood criterion. Hence, the utility of internal validation is limited.

External validation means comparing clustering solutions to ground truth data. This approach uses statistical indexes or distance/similarity functions to calculate how similar a clustering solution is to a given ground truth solution. It is important that the external ground truth is independent of the data used for clustering. For example, evaluating a clustering solution constructed from co-citation data with a criterion calculated from direct citation information would lead to skewed and uninterpretable results, because high criterion values are a foregone conclusion. On the other hand, evaluating a co-citation based solution using a criterion calculated from keyword co-occurrence data would satisfy the independence condition.

A common source of ground truth data is data from established classification systems or other knowledge organization systems such as thesauri. Palchykov, Gemmetto, Boyarsky, & Garlaschelli (2016) compare the clustering solution of papers from the preprint service arXiv on the basis of similarity of concepts with the arXiv topic classes (13 subclasses of physics chosen by the authors). Zhang et al. (2018) select sample article records from 10 Web of Science Subject Categories. After clustering the records, the Categories are used as the ground truth. Ahlgren, Chen, Colliander, & van Eck (2020) evaluate various publication similarity estimation methods by how well the clustering solutions obtained from them match with aggregated data of Medical Subject Headings. Other types of ground truth data have also been applied. Klavans & Boyack (2017) take the sets of references of articles with at least one hundred references as proxies for topics. This method is also used by Sjögårde & Ahlgren (2018). Boyack & Klavans (2010) use sets of papers acknowledging the same funding grant as topic proxies. Others have used the publications of journals or conference series as constituting the ground truth classes (or as proxies of them). For instance, Sjögårde & Ahlgren (2020) follow this method for the construction of a baseline classification of "specialties", the grouping structure hierarchically one level above the "topics" of Sjögårde & Ahlgren (2018). Validation with external data containing structural information is also the strategy used in the present study.

A more rarely used validation strategy is expert validation. Experienced field experts analyze the clustering results and judge whether or how much they are in agreement with their expectations. Šubelj et al. (2016) apply this evaluation approach to clustering solutions of a dataset of publications in library and information science that covers scientometric topics. Zhang et al. (2018) requested experts in bibliometrics to assess the clustering solution of their proposed method on a set of papers in three journals covering bibliometric research. They rated eight clusters based on their extracted terms on their coherence, distinctiveness and significance. In a second round, a different method was pursued. Experts were given the forty extracted terms and were asked to group them into coherent clusters. These expert clusters were then used as the standard against which their algorithmic solution was benchmarked. Such validation exercises are very labor-intensive and reflect the subjective views of the chosen experts. There seems to be no published standard protocol for expert validation of clustering solutions.

2 Data

As a preliminary to introducing the data used in this study we should like to address the following points. An important consideration in the application of external data for clustering validation is that it is not necessary that the validation data covers all elements of the dataset. If validation data is only available for a subset of the data one can restrict the evaluation of clustering solutions to the same subset. In such a case, one must take care that the subset is not affected by selection bias. A second consideration is that the assumption that a single external classification is the one true ground truth can not be justified. Multiple perspectives on how to classify research may exist side-by-side. A single classification system may be too limited to accomodate for these. Palchykov et al. (2016) point out that research classification systems are commonly designed to reflect the structure of the phenomena under study. In doing so, they may be incapable of taking into account the diversity of methods used in studying the same phenomenon. Other criteria might be orientation towards fundamental insight (basic science) versus immediate use in practice (applied research) or the categorization into theoretical, methodological and empirical research. For this reason, we employ a number of validation datasets which are both independent of the clustered data and of each other, all containing some signal of the structure of science as perceived by researchers, without preferring any single one dataset over the others. Instead we expect them to mutually complement each other. Furthermore, we use the term "validation data" instead of "ground truth", as we consider it more appropriate. The most important consideration in our choice of validation datasets was that these validation datasets need to be logically independent of the data and method used for clustering. Concerning clustering method, it is misguided to evaluate clustering solutions by using criterion metrics for clustering quality that were themselves the optimized function in the clustering process. It would also be misguided to evaluate solutions using data that was already the input for the clustering procedure (cf. Waltman, Boyack, Colavizza, & van Eck (2020), section 2.4). For example, comparing a clustering solution based on text similarity with one based on direct citations against a gold standard based on, say, bibliographic coupling would inevitably advantage the direct citation clustering as both direct citation and bibliographic coupling use the same citation network and therefore closely related similarity signals. This concern has led us to not adopt the method suggested in Klavans & Boyack (2017) of using the references of papers with more than 100 references as gold standard groups. Adopting this method would advantage citation network-based clustering solutions over those based on text similarity because the citation-based methods directly incorporate the validation data structure. It would be different if those validation documents and their references were removed from the citation network for creating the clustering solutions, as this would restore the necessary independence.

The second consideration for choosing validation data is that we would like to get data that as closely as possible derives from the perceptions of researchers active in the field under analysis. In the subsections on the four validation datasets this issue will be taken up individually for each dataset.

2.1 Topic Extraction Challenge dataset and clustering solutions

The Topic Extraction Challenge *Astro* bibliometric dataset was obtained from Clarivate Analytics, the company which produces the Web of Science. The *Astro* dataset was derived from Web of Science data, for a description see Gläser et al. (2017). In summary, papers (articles, letters, and proceedings papers) from journals that are classified in the Subject Category *Astronomy and Astrophysics* in Web of Science, published from 2003 to 2010, were collected. Not all journals are exclusively about astronomy and astrophysics (for example *Physical Review D* and *Comptes Rendus Physique*).

The submitted clustering solutions of this dataset were downloaded from the project website[2]. We excluded the solution `hd` of Havemann, Gläser, & Heinz (2017) because it is of a qualitatively different nature, namely consisting of overlapping communities, and cannot be meaningfully compared with the other solutions by our validation approach. In this solution, unlike the others, one publication can be assigned to more than one cluster. While the solution contains cluster assignments to 101,831 publication records, just six records are assigned to only one cluster. Moreover, most assignments are with the maximum strength (1.0). If one removes all but the highest scored cluster assignments for each record one is still left with three different cluster assignments for each record on average, all tied at strength=1.0. It is hence not reasonably possible to reduce the overlapping communities of `hd` to a hard clustering, as there are too many possibilities.

All seven other solutions of the first round of the Topic Extraction Challenge assign only a single cluster to each publication record. Solution `c` by van Eck & Waltman (2017) uses direct citation as the data model and applies the authors' clustering algorithm, which optimizes a variant of the modularity function. Glänzel & Thijs (2017) contribute solution `eb`, using bibliographic coupling, and solution `en`, using a hybrid of bibliographic coupling and natural language processing. Both of these are clustered with the Louvain community detection algorithm, an algorithm optimizing the modularity function. Solutions `ok` and `ol` by Wang & Koopman (2017) are obtained by using Louvain and k-means clustering of similarity data obtained from a semantic matrix which incorporates signals from various document features such as words, references,

[2] http://141.20.126.171/solutions.html

authors and journals. The solution `sr` (Boyack, 2017) uses the same clustering algorithm and data model (direct citations) as solution `c` but instead of using the citation relations from the *Astro* corpus it experiments with using global citation relations, that is, including citation links from outside the primary dataset. While the other solutions have on the order of 10 to 30 clusters, `sr` consists of 555 clusters. Solution `u` uses the Infomap clustering algorithm in two iterations on a direct citation data model (Velden et al., 2017). Some more descriptive data on the solutions is presented in Table 4.

Xu et al. (2018) have also used the *Astro* corpus for testing a mixed-membership stochastic blockmodel clustering procedure which produced overlapping cluster assignments. They experimented with three citation-based data models: direct citation, bibliographic coupling, and co-citation. While they also produced solutions in which the number of clusters was left unparametrized, they published three clustering solutions in which the number of clusters was fixed to 113 to enable direct comparison with solution `hd`. We downloaded these solutions[3] and simplified them such that each item is assigned only to one cluster, the one with the highest strength value, the same way the authors did. Contrary to the `hd` solution, this was straightforward in this case, because there was almost always one cluster with a clear highest value and very few ties. We use the following abbreviations for these solutions:

- `mb`: `m`ixed-membership stochastic blockmodel, `b`ibliographic coupling
- `mc`: `m`ixed-membership stochastic blockmodel, `c`o-citation
- `md`: `m`ixed-membership stochastic blockmodel, `d`irect citation

2.2 NSF grant linkage data

We now turn to the four validation datasets. Many research projects are financed through competitively acquired grants from research funding organizations which typically award such grants based on committee peer review of submitted project applications. When reporting the results of the funded research project in publications, the funding sources are by convention listed in the acknowledgements section of those publications. Typically, the name of the funding organization and the project ID are documented. It is therefore natural to expect all publications that acknowledge funding from the same grant, that is, the same project, to belong to the same research topic, given that most regurlar research project are associated with one topic. This reasoning has first been put forward and successfully exploited for clustering validation by Boyack & Klavans (2010) and Boyack et al. (2011), who refer to this data as *grant linkage data*.

We make the following two simplifying assumptions about the relationship between grants and topics. (1) Grant-financed projects are concerned with one specific research topic. The reality can, depending on one's adopted definition of a topic, be more complicated. But we assume that one grant is at least a reasonable proxy for one topic. (2) Conversely, a topic does not consist only of a single grant's publications. From this it follows that if two publications have funding from different grants it does not mean that they belong to different topics. Furthermore, publications often report funding from more than one grant. This does not mean such a publication belongs to several topics but rather the acknowledged grants belong to one topic. Grant-linkage data represents researchers' views of topical structures as grant applicants formulate the research topic of the grant as a cohesive whole and convince peer committees of fellow researchers of the topic's relevance and the soundness of their proposal.

We constructed article-grant linkage validation data for the *Astro* dataset by the following procedure. We

[3] http://54xushuo.net/wiki/lib/exe/fetch.php?media=resources:datasets:xlza_2018.zip

chose the US National Science Foundation (NSF) as source of funding data, as it is a large funding agency with a significant program for astronomy and it documents both grant information and the publications that resulted from the granted projects and makes this information easily accessible. From the NSF website's advanced award search interface we selected grant data from the "AST Division of Astronomical Sciences" section with award start dates from 2000-01-01, which is three years before the coverage of the *Astro* corpus begins, to 2010-12-31. We downloaded the results as a CSV file with 2359 grant records. This dataset contains the award IDs but not the metadata of the publications that resulted from granted projects. We used an `R` script to request the publication metadata for each of the grants from the NSF Awards API using the award IDs. In particular, we collected the API result set records fields `id`, `title`, `fundProgramName`, `publicationResearch`, `publicationConference`. From the returned XML data we extracted award numbers (`id`) and various bibliographic fields of each publication record. The data contain a URL to the Web of Science (WoS) platform page of the publications and this URL includes the unique WoS ID. We extracted only this ID from the URL, as it can be used directly to match with the *Astro* dataset `UT` field. After matching with the *Astro* dataset `UT` field and removing records of grants having only a single matched publication record, the final validation dataset consists of 477 grants, representing groups of closely associated publications, with a total of 4,271 article-grant linkage relations. The sizes of the publication groups range from 2 to 98 publications per grant. There are a number of publications with more than one grant, as is shown in Table 1.

Table 1: Incidence of multi-grant papers

grants	publications
1	4275
2	687
3	170
4	25
5	9
6	2

2.3 Special issues data

Scientific journals occasionally publish topical special issues. Special issues are often announced by a public call for papers on a defined topic of particular current interest. In such a call for papers the special issue's topic and the expected content of submissions is outlined. Frequently, special issues are handled by invited guest editors who are experts in the topic. As topical special issues are created at the initiative of researchers they closely reflect their perceptions of what constitutes topics and which publications belong to a topic.

We therefore use special issues as another proxy for the topics exptected to be found by clustering procedures. Similar to the situation for grant-linkage data, we do not expect one topic to map to exactly one special issue. Special issues of various journals from different periods should be in the same topic cluster if they cover the same topic. However, all articles in a single special issue are expected to be about the same topic, so they should all be in one topic cluster.

For the source journals in the *Astro* dataset we checked the journals' websites for the issues of the covered time period and collected data on all topical special issues. Special issues not related to specific topics, namely those for general astronomical conferences, were not considered. The collected special issues data contains all *IAU Symposia* series issues, except one issue titled "HIGHLIGHTS OF ASTRONOMY, VOL 13" in the *Astro* dataset, published 2005. The data also includes all issues of *Advances in Space Research* with specific issue titles in the dataset. Both of these are serials containing proceedings papers from conferences, colloquia and workshops on specific topics by the International Astronomical Union (IAU) and the International Council for Science Committee on Space Research, respectively. *Highlights of Astronomy* was a separate series of publications of the IAU. We found that there is only one publication record in the *Astro* dataset for the journal *Experimental Astronomy*, volume 26, issue 1-3 (special issue on 400 Years of Astronomical Telescopes) although there should be 14 and have therefore not included this issue. The dataset also contains only three out of 13 items of *General Relativity and Gravitation*, volume 42, issue 9 (special issue on Gravitational Lensing) and 3 out of 10 items of *New Astronomy Reviews*, volume 52, issue 6 (special issue on Active Galactic Nuclei at the Highest Angular Resolution: Theory and Observations) but we have kept these records.

This data collection resulted in a final dataset with a total of 423 special issues from 24 different journals with 11,329 publications in total. The groups range in size between 3 and 220 items. There are 26.8 items per special issue on average. The data also include some non-astronomical groups from journals which are only partially about astronomy. It contains, for example, special issues on "Aircraft trailing vortices", "Carbon nanotube electronics" from *Comptes Rendus Physique* and "The Earth's Deep Interior" from *Geophysical and Astrophysical Fluid Dynamics.*

2.4 Journal topic classifications data

Some journals maintain their own custom classification systems by which each article is classified into a single class in order to provide a degree of knowledge organization for readers. Such article level classifications are directly based on active researchers' perceptions of topical structures because journal editors – experienced researchers themselves – are responsible for introducing and maintaining such classification systems for their journals and because authors and editors are responsible for classifying accepted papers appropriately.

The papers in a journal issue might be grouped by these classes under their headings or the classes might be labels for each paper. For our purposes this is irrelevant. What matters is that each article has one of a finite set of pre-existing classes assigned to it.

We argue that such journal-specific classifications of general disciplinary journals are another useful proxy for topics. These classifications were created by journal editors for the purpose of organizing their journals' contents for ease of use by scientist readers. In a topic clustering solution the articles of the same classes should therefore be concentrated into one or a few clusters rather than dispersed over many clusters.

We have checked the journal websites of the *Astro* corpus journals for the covered time period for such classifications. Four of the journals in the dataset maintained such classifications, namely *Publications of the Astronomical Society of the Pacific*, *Astronomy & Astrophysics*, *Chinese Journal of Astronomy and Astrophysics* and *Kinematics and Physics of Celestial Bodies.* None of these journals use the same classification system. We web-scraped and manually collected the journal topic classification data for all articles in these journals for the time period covered by the data and matched them to the *Astro* dataset records by DOIs or by manual look-up. We cleaned the data by removing classes that are not about a topic, but about publication types, such as "invited reviews" and "opening remarks" and unifying obvious small errors and inconsistencies such as missing definite article "the" in the class label or changing class label from "Ttars" to "Stars" among others.

Table 2 gives an overview of the number of classes and papers for the four astronomy journals used for clustering validation. The complete enumeration of all used sections is given in Table 9 in the Appendix.

Table 2: Summary of journal topic classification data

journal	topic classes	papers
Astronomy & Astrophysics	16	12,502
Chinese Journal of Astronomy and Astrophysics	24	565
Kinematics and Physics of Celestial Bodies	13	105
Publications of the Astronomical Society of the Pacific	14	855

As for the validity of these classification systems, we assume that they are at least reasonably good representations or proxies of the intellectual structure of the field. However, one has to keep in mind that these classification systems are themselves not perfect ground truths. Waltman et al. (2020) remark on this issue

> There is no perfect classification of publications that can be used to evaluate the accuracy of different clustering solutions. For instance, suppose we study the degree to which a clustering solution resembles an existing classification of publications [. . .] The difficulty then is that it is not clear how discrepancies between the clustering solution and the existing classification should be interpreted. Such discrepancies could indicate shortcomings of the clustering solution, but they could equally well reflect problems of the existing classification.

2.5 ADS *also read* usage data

The SAO/NASA Astrophysics Data System (ADS) is an online bibliographic and bibliometric database covering astronomy and astrophysics publications (Kurtz et al., 2005; Kurtz & Henneken, 2014). ADS offers several second order query capabilities. These take as arguments lists of query results and return new result lists based on the input list's attributes (Kurtz, Eichhorn, Accomazzi, Grant, & Murray, 2002). One of these functions, implemented as the `trending()` operator, enables users to find records of papers that were read by ADS users who have read papers on the input list. On each article's abstract page, ADS also shows papers that readers of that specific paper have also read – the result of the `trending()` function for the currently viewed document as its input. ADS initially called this feature *also read* but it is now referred to as *co-reads*. We will use the older name here, because it was in use at the time of our data collection. These data are constructed from usage log files of the past 90 days, filtered for users who show the usage patterns of active researchers. As ADS is used heavily by research-active astronomers, the logged activity patterns closely reflect search and usage of literature by professionals in the discipline. The utility of usage log data for science mapping has been pointed out and exploited successfully by Bollen et al. (2009) who used aggregated clickstream data from several different scientific literature portals to construct journal maps of all sciences. The authors list as advantages of usage data over citations that the usage data is more up-to-date, has a far higher number of observations per item and includes utilization of persons other than publishing researchers.

We collected *also read* information for publication records in the *Astro* dataset by web-scraping in February 2018. At that time, the actual values for the number of common readers of a record and its *also read* records was displayed in the web interface (cf., Kurtz et al., 2002, p. 244, Fig. 6), which is no longer the case. We obtained ADS data by the following process. We searched for all *Astro* records' DOIs in ADS and saved their ADS-internal identifiers, the `bibkey`, if an exact match was found. For those *Astro* records for which we could not identify a `bibkey` by DOI we also searched for the combination of author names, publication year, and title in ADS and saved the `bibkey` if there was a result that was an exact match (match score of 1.000). With all found *Astro* `bibkeys`, we queried ADS and saved the `bibkeys` of the *also read* publication records and their read counts. In the result data, we translated all `bibkeys` to the *Astro* UT identifiers and removed all cases where there was no *Astro* record for a `bibkey`. These are records in the *also read* lists of *Astro* papers which are themselves not part of the *Astro* dataset. The maximum co-usage score in the dataset is 92. The dataset consists of 2,168,578 record pairs of 57,268 unique items. The average co-usage count is 1.3.

Unlike the datasets described above, this one does not contain any groups of items, only pairs. The *also read* count can be seen as a kind of similarity measure based on user behavior but not as a topic classification. The higher the number of user sessions in which two items were used, the more similar the items are expected to be. While the data would be more directly useful for validating similarity calculation approaches, it can also be used for cluster validation because highly similar items should be members of the same cluster. A solution that better groups co-used pairs into common clusters is better able to represent the co-usage structure of the data.

2.6 Remarks on the data

In total, these four validation datasets contain information on 85302 unique *Astro* items, a coverage of about 76.4%. Co-readership, special issues and journal sections data have so far never been used in the evaluation of bibliometric publication clustering solutions and are therefore a novel contribution of this study. We would like to point out that the datasets presented above do not exhaust the possibilities of good external evaluation data for the *Astro* corpus. On the contrary, in particular for astronomy and astrophysics, there are two knowledge organization systems which have been in use at several journals each in the covered time period. Namely, the Physics and Astronomy Classification System (PACS), used for instance by *Physical Review D*, *Gravitation and Cosmology*, *International Journal of Modern Physics D*, *Kinematics and Physics of Celestial Bodies*, *New Astronomy*, and *Il Nuovo Cimento C*, and the Astronomical Subject Keywords, used by *Astrophysical Journal*, *Astronomical Journal*, *Astronomy & Astrophysics*, *MNRA*, *Publications of the Astronomical Society of Japan*, *Publications of the Astronomical Society of the Pacific*, and *Revista Mexicana de Astronomia y Astrofisica*. Both systems are now retired and superseded by new systems but were in active use in the period that the *Astro* dataset covers. We chose not to include them in this study as it was not feasible to acquire the complete and clean data with proportionate effort. Nevertheless, data derived from these systems would be an obvious worthwhile extension to our data.

The application of external validation ("ground truth") data for verifying clustering solutions has been critized by the contention that node metadata and ground truth are not equivalent (Peel, Larremore, & Clauset, 2017). We therefore stress that the external data we use are not mere node metadata. The reasons for the appropriateness of the datasets as *proxies* for topical structures have been explicitly articulated in the respective sections above. We do not claim that any of the datasets is a ground truth in the strict literal sense of being the factually known correct grouping of items – such a thing does not exist outside of synthetic datasets. Rather, we employ the term validation data (instead of ground truth) for a dataset that has a structure affected by the latent structure of the analyzed data and is therefore a valid proxy for the structural properties of the data. We also agree with the notion that there is no single correct verification dataset for real-world data. This is why we intentionally choose a number of data sources that are both independent from the data used to cluster the *Astro* dataset and independent of each other. If there is a latent structure in the data which can be uncovered by clustering, we assume that all partial validation datasets are affected by it and capture some part of this structure. In combination, the datasets should account for more of the structural information than any single dataset would, i.e. the datasets are not redundant with respect to the validation carried out here. We are therefore confident that this study can further contribute to the understanding of existing clustering approaches and is to some degree objective.

3 Methods

We evaluate each of the *Astro* clustering solutions by how much they structurally align with the four validation datasets introduced above. These datasets are used as they are and are not further clustered. We are therefore not comparing clusterings with clusterings but clusterings with different kinds of grouping structures. While further clustering the validation datasets would be possible, for example one might wish to agregate all NSF grants which appear to be about a single topic, this would mean that any such clustering, including the measurement of group similarities, would itself first have to be evaluated. Here we opt to simply use the grouping structure of the validation data as it is, because it is already appropriate for the purpose.

To measure the agreement between the external partial validation datasets, a comparison measure is needed. Many different external clustering evaluation metrics have been proposed and discussed in the methods literature. These are generally intended to be used in situations in which clustering solutions are being compared to a reference partition of a validation dataset. For many measures, it is necessary that both compared solutions are, in terms of mathematical set theory, *partitions of the set* of objects, that is, subsets such that each element is member of exactly one subset. The validation data we have presented is not in such a structure and cannot be transformed into it, which rules out the application of such evaluation metrics. The ADS *also read* data is structured as pairs and the three other datasets can be seen as subgroups of topic clusters which are not reconciled into a single unified partition structure. That leaves pairs-based measures, the most well-known example of which is the Rand Index. For this type of measure, it is of interest whether given objects pairs that are linked in the validation data are assigned to the same cluster in the clustering results.

An appropriate evaluation measure for the validation data at hand is the true positive rate (TPR), the ratio of true positive pairs to positive pairs. In the present case, given the publication items pairs in a validation dataset, the TPR is the share of these pairs for which both items are in the same cluster in a solution.

There is an additional problem to take into account. As it stands, this index would be biased towards clustering solutions with a low number of clusters and large cluster sizes. We correct for this by making an adjustment for chance (Meila, 2015, p. 624; Vinh, Epps, & Bailey, 2010, p. 2843ff). Five random clustering solutions are generated with the same cluster sizes as each tested solution and TPR is calculated for these randomized solutions. Then the TPR for the actual clustering solution is rescaled by setting the average of the random clusterings' TPR as the zero point according to the formula

$$TPR_{adjusted} = \frac{TPR_{observed} - \overline{TPR}_{random}}{1 - \overline{TPR}_{random}},$$

where $\overline{TPR}_{random}$ indicates the arithmetic mean of the five randomly generated TPR values. This removes the confounding effect of the different size distributions of the clustering solution so that the different solutions can be compared. We note that the values for the five random TPR per solution and validation dataset in all cases have very similar values.

Note that because the validation datasets only partially cover the data and because they are not partitions, we have no way to measure the purity of clusters, that is, to which degree clusters are composed of items of only one group of the validation dataset.

4 Results

4.1 A plausibility check for document similarity measures

As it is perhaps not immediately convincing that the *also read* scores introduced above are indeed useful for validating or identifying topics, we briefly check if they can serve as a reasonable document similarity measure. The values of a good document similarity measure should accord with our expectations of what we know are typically more or less similar publications. We can define classes of publications for which we expect similarity scores of randomly chosen members to be of smaller of greater average magnitude. We provide one initial sketch of such classes, ordered by expected average member document similarity. While the details are subject to debate and our list is certainly incomplete, the general idea is very simple and intuitive.

We posit that average similarity values should increase (and relative variability of similarity values decrease) in this order:

1. random papers; should have almost zero similarity
2. papers in the same general science journal covering different disciplines (e.g., *Science*, *Nature*, *PNAS*, *PLoS ONE*)
3. - papers in the same general disciplinary journal (such as *Journal of the American Chemical Society*, *New England Journal of Medicine*, *Astronomy & Astrophysics*)
   - books in a book series on a discipline
4. - papers in the same specialist journal (most other journals, e.g., *Solar Physics*, *Scientometrics*)
   - topical sections (see section 2.4) or topical special issues in general disciplinary journals (e.g. *Transactions* journals of the Royal Society)
   - chapters in an edited book on a topic
   - books in book series on a given topic
5. papers with direct citation, with similarities between a given paper and its references being correlated to the number of in-text citations of the references in the citing document
6. - papers in special issues in specialist journals, (see section 2.3)
   - papers originating from the same research project, identifiable by having the same grant acknowledged (see section 2.2)

Note that all the above are intentionally chosen as to be in principle readily available in existing databases. We would also expect certain degrees of similarity of publications of:

- the works of a team/lab/group
- the works of a department/institute
- publications collated as a cumulative PhD thesis
- PhD theses supervised by one professor

Table 3: Average ADS also read scores for different classes of documents

data	observations	non-zero obs.	average also read score * 100	coefficient of variation
random papers (10% sample)	63,417,623	11,609	0.02	100.76
journals				
Astrophysical Journal	191,717,571	90,936	0.07	73.44
Astronomy & Astrophysics	107,567,778	83,884	0.09	43.42
MNRAS	67,262,601	78,104	0.15	42.06
Astronomical Journal	5,819,166	5,926	0.20	51.05
grouping structures				
journal sections	9,013,340	21,928	0.30	25.16
special issues	285,202	629	0.34	27.92
NSF grants	44,191	1,356	6.40	8.08

However, it seems far less clear where to place them in the above scheme.

For purposes of validating the *also read* data, we have computed the average value and coefficient of variation of the *also read* score for the classes of documents collected for this study, namely publications acknowledging the same grant, publications from the same special issue, and publications from the same topic section of a journal. We also include the values for papers of four selected journals. Most journals have either no or very few non-zero observations and hence zero or very low average *also read* score. This is true mostly for the smaller journals, for which co-usage of two papers from the same journal by ADS users was negligible at the time of our data acquisition. Thus it should be kept in mind that the analyzed journals are the large ones in the field and their values are not representative for all journals in the dataset. As a baseline, we also calculated the average *also read* value of random papers in the *Astro* dataset. If no *also read* score for a pair was present in the data, a value of zero was used in all four classes of papers. The results are shown in Table 3. In line with the predictions, the average *also read* values of random documents is the smallest. The co-usage values for papers from the large selected journals are higher. The values of documents from the same journal section and from the same special issues are nearly equal and by a fector of ten greater than random papers. Note that the coefficient of variation is far smaller than that for random papers. This indicates that the similarity values of documents in these similarity classes tend to concentrate around the average more than those of randomly chosen documents. Finally, the average similarity of documents funded by the same NSF grant is an order of magnitude higher than that of documents from the same journal topic section or the same special issue. Our assumption that the similarities of papers in a special issue are as concentrated on a topic as those of a single granted project was probably not justified. The relative variability of *also read* similarity values of publications of the same granted project is also much lower than that of the other two classes of publications. This also supports the notion that granted projects are topically more narrow than special issues and journal sections. Note also the increase in the proportion of non-zero observations as the presumed similarity increases. These results indicate that the *also read* values are, at least on a large scale, in line with expectations of a reasonable document similarity measure.

4.2 Clustering solutions validation

Table 4: Solutions description and results

					validation results rank			
solution	data model	clustering algorithm	clusters	reference	NSF grant linkage	special issues	journal topics	ADS also read
c	direct citation	SLMA	22	van Eck & Waltman (2017)	2	4	5	3
eb	bibliographic coupling	Louvain	13	Glänzel & Thijs (2017)	6	1	4	7
en	bibliographic coupling + NLP	Louvain	11	Glänzel & Thijs (2017)	7	2	3	4
mb	bibliographic coupling	MMSB	113	Xu et al. (2018)	10	9	9	9
mc	co-citation	MMSB	113	Xu et al. (2018)	8	8	8	8
md	direct citation	MMSB	113	Xu et al. (2018)	9	10	10	10
ok	semantic matrix	k-means	30	Wang & Koopman (2017)	4	7	6	5
ol	semantic matrix	Louvain	32	Wang & Koopman (2017)	5	6	7	6
sr	global direct citation	SLMA	555	Boyack (2017)	1	5	1	1
u	direct citation	Infomap	22	Velden et al. (2017)	3	3	2	2

Note:

NLP: Natural Language Processing; SLMA: Smart Local Moving Algorithm; MMSB: Mixed-Membership Stochastic Blockmodel

In this section we present the adjusted true positive rate results for the four validation datasets. The results are summarized, together with the characteristics of the clustering procedures, in Table 4 and visualized in Figure 1. The four individual dataset results tables also include the details on the calculation of the values for each solution. Table 5 displays the results for the calculation of the adjusted true positive rate for the NSF grant linkage data. Solution `sr` achieves the best result, followed at some distance by `c` and `u`.

Table 5: Results NSF grant linkage data

solution	true positive pairs	pairs	TPR	avg. random TPR	adjusted TPR
c	26,220	44,097	0.595	0.060	**0.569**
eb	19,969	43,803	0.456	0.082	**0.407**
en	20,709	43,768	0.473	0.112	**0.407**
mb	9,929	43,030	0.231	0.010	**0.223**
mc	8,282	34,800	0.238	0.010	**0.230**
md	9,844	42,539	0.231	0.010	**0.224**
ok	19,178	40,391	0.475	0.040	**0.453**
ol	20,344	44,191	0.460	0.040	**0.438**
sr	30,130	41,226	0.731	0.152	**0.683**
u	25,593	44,097	0.580	0.080	**0.544**

The results for the validation with the special issues dataset, presented in Table 6, show that the bibliographic coupling solutions `eb` and `en` perform best with values of about 0.5 while the other solutions obtain values at lower levels.

Table 6: Results special issues data

solution	true positive pairs	pairs	TPR	avg. random TPR	adjusted TPR
c	81,173	171,138	0.474	0.060	**0.441**
eb	135,045	244,751	0.552	0.082	**0.512**
en	151,087	269,239	0.561	0.110	**0.507**
mb	20,070	127,846	0.157	0.010	**0.148**
mc	20,618	75,120	0.274	0.010	**0.267**
md	18,707	121,623	0.154	0.010	**0.145**
ok	117,065	276,630	0.423	0.040	**0.399**
ol	123,881	285,202	0.434	0.040	**0.411**
sr	102,645	198,826	0.516	0.150	**0.431**
u	85,909	171,138	0.502	0.080	**0.459**

The results for the journal topic classifications systems validation dataset are presented in Table 7. Here also, solution `sr` has the best result by a large margin, followed by solution `u`.

In the results for the ADS *also read* data, shown in Table 8, the solution `sr` is again the one with the best score, followed by `u`.

In summary, solution `sr` achieves the best results in three of four tests, while `eb` and `en` do well in the special issues dataset[4]. It can be seen that the solutions which use similar clustering approaches, namely (`c`, `u`), (`eb`,

[4] The reason for this result could be that these two solutions, `eb` and `en`, unlike the other ones, do not rely on direct citation, which is likely to be relatively rare between papers of a single special issue, but on bibliographic coupling and NLP-enhanced text similarity.

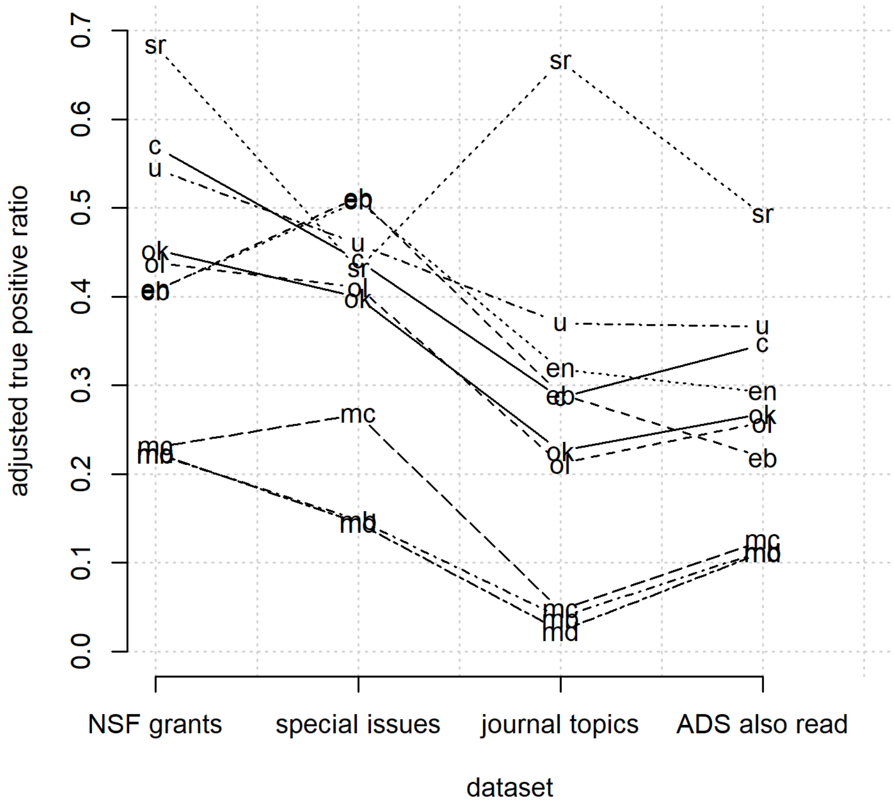

Figure 1: Comparison of TPR scores of solutions across validation datasets

Table 7: Results journal topic classification data

solution	true positive pairs	pairs	TPR	avg. random TPR	adjusted TPR
c	2,928,659	8,897,742	0.329	0.060	**0.286**
eb	3,029,053	8,744,104	0.346	0.080	**0.290**
en	3,440,435	8,714,308	0.395	0.112	**0.318**
mb	400,214	8,395,017	0.048	0.010	**0.038**
mc	375,888	6,634,123	0.057	0.010	**0.047**
md	267,396	8,163,134	0.033	0.010	**0.023**
ok	2,073,438	8,067,206	0.257	0.040	**0.226**
ol	2,194,111	9,013,340	0.243	0.040	**0.212**
sr	6,269,709	8,770,170	0.715	0.150	**0.665**
u	3,742,662	8,897,742	0.421	0.080	**0.370**

Table 8: Results ADS also read data

solution	true positive pairs	pairs	weighted TPR	avg. random weighted TPR	adjusted TPR
c	877,900	2,146,364	0.386	0.062	**0.346**
eb	602,802	2,106,634	0.285	0.084	**0.219**
en	776,690	2,101,242	0.372	0.113	**0.292**
mb	248,494	2,063,912	0.121	0.009	**0.113**
mc	224,678	1,674,360	0.133	0.010	**0.125**
md	238,068	2,060,676	0.119	0.009	**0.112**
ok	566,806	1,837,136	0.295	0.037	**0.268**
ol	660,554	2,168,578	0.290	0.044	**0.258**
sr	1,196,310	2,046,920	0.569	0.152	**0.492**
u	952,528	2,146,364	0.417	0.081	**0.366**

en), (ok, ol), and (mb, mc, md) tend to have similar results. Solutions eb and en only deviate for dataset ADS *also read*. There, en performs better, which is the solution that in addition to bibliographic coupling also integrates sophisticated NLP-based lexical similarity. Solutions c and u deviate for dataset journal topics, while solutions ok and ol score similarly on all four datasets. The stochastic blockmodel solutions results mb/c/d only differ in the special issues dataset, in which the co-citation solution performs better.

Furthermore, the order of the solutions' results values between the three datasets NSF grants, journal topics, and ADS *also read*, are consistent to some degree while the special issues dataset is clearly different, judging from the solution results pattern. The item groups of the datasets journal sections and special issues are by their nature constrained to be within specific journals. Solutions ol and ok explicitly use the information on which journals the publications belong to as a feature and could therefore be advantaged in these parts of the validation. However, these are the two datasets in which ol and ok have their worst results. One possible explanation might be that the journal by itself is not a feature with a good discrimination ability. Koopman, Wang, & Scharnhorst (2017) also remark that they used the WoS standard author abbreviation and that they did not disambiguate author names which could also have led to conflation of the discriminatory signal of individual authors' topical specializations.

The overall rather poor results of the three solutions using the mixed-membership stochastic blockmodel are in line with the conclusions of Xu et al. (2018). The authors found that the solutions they obtained were not satisfactory based on a comparison with solution `hd` of Havemann et al. (2017) and the unusally even distribution of cluster sizes. The results obtained in the present study corroborate this finding by showing that, irrespective of the data model, the solutions evince validation scores below the level of other clustering algorithms' solutions.

The compared solutions are affected differently by the adjustment for chanceas can be appreciated by comparing the columns TPR and adjusted TPR and the values in the average random TPR column in the results tables. Solution `sr` has the highest values for the TPR obtained by randomly constructed groupings with the same numbers and sizes of clusters. This solution is dominated by a few very large clusters. Random clusterings with the same characteristics achieve TPR values of about 0.15 on these validation datasets, that is, 15% of the item pairs would have both items in the same (random) cluster. This is on par with the results of the poorer performances of some solutions in some datasets. Yet all other solutions' values are also attenuated by the correction of chance agreement, just less conspicuously so. This demonstrates well how important it is to adjust for chance when comparing solutions of different size distributions.

To provide a more detailed analysis of the results, we have calculated the average true positive rates of the four datasets for each solution on a per-cluster basis. The three clusters with the best and worst scores, restricted to clusters for which values for all four datasets could be calculated, are presented in Table 10 in the Appendix. Due to the large amount of data, we have limited the analysis to these selected results. We find that several solutions are able to construct high-quality clusters for specific topics (cluster labels are from Velden, Boyack, et al. (2017), Table 4). We only have cluster labels for the first round of papers of the Topic Extraction Challenge, so clusters from the three solutions of Xu et al. (2018) are not taken into account here. The high quality clusters are as follows.

- on solar activity:
  - `c 3`, "solar, coronal, active region, cme, flare, magnetic field, sunspot, mass ejections, quiet sun, chromosphere"
  - `eb 4`, "solar, magnetic field, coronal mass, solar activity, plasma, active region, ionospheric, cme, flare, sunspot"
  - `en 11`, "magnetic field, solar wind, plasma, coronal mass, ionospheric, active region, waves, solar activity, field lines, reconnection"
  - `ok 14`, "coronal, active region, solar, flare, magnetic flux, cme, quiet sun, chromosphere, mass ejections, hinode"
  - `ol 8`, "coronal mass, solar, cme, active region, flare, mass ejections, magnetic field, magnetic reconnection, chromosphere, transition region"
  - `sr 400`, "solar, coronal, active region, cme, flare, sunspot, mass ejections, magnetic flux, quiet sun, transition region"
  - `u 4`, "solar, coronal mass, active region, cme, flare, magnetic field, mass ejections, sunspot, quiet sun, chromosphere"
- on galaxies/active galactic nuclei:
  - `c 1`, "galaxies, redshift, star formation, sample, active galactic, agn, gas, galaxy clusters, digital sky, sloan digital"
  - `en 8`, "star formation, galaxies, formation rate, digital sky, sloan digital, sky survey, molecular gas, gas, sample, h ii"
  - `sr 126`, "galaxies, redshift, star formation, clusters, sample, active galactic, agn, sloan digital,

digital sky, halo"
  - u 1, "galaxies, redshift, active galactic, agn, star formation, galactic nuclei, quasar, sample, gas, galaxy clusters"
- on stars:
  - eb 1, "star, radial velocity, planet, orbital period, hd, transit, binary, eclipsing binary, main sequence, white dwarf"
  - en 1, "stars, main sequence, radial velocity, giant branch, photometry, fe h, binary, red giant, asymptotic giant, mass loss"
  - ok 2, "transit, star, eclipsing binary, radial velocity, hd, planet, corot, photometric, main sequence, type stars"
  - sr 17, "star, main sequence, binary, light curve, gamma ray, white dwarf, neutron star, emission, low mass, x ray"
  - u 3, "star, planets, hd, main sequence, brown dwarfs, radial velocity, planet formation, transit, type stars, extrasolar planets"

On the other hand, for the topic of cosmic microwave background, one solution has a high quality cluster (ol 18), while two solutions have low quality clusters (c 2, eb 11).

5 Discussion

We have conducted an independent validation study on ten topic clustering solutions for a publication records benchmark dataset from astronomy and astrophysics. We collected external validation data and introduced three novel types of validation data. The first main contribution of the paper is establishing within-journal topic categories, special issues, and scholarly bibliographic platform co-read data as new sources of topical grouping validation data for scientific document clustering. The second contribution is using these datasets and the clustering approach using grant linkages (Boyack & Klavans, 2010) to validate clustering solutions of the common publication dataset contributed by various research groups. A limitation of our validation datasets is that they are not exhaustive. In particular, datasets of assignments of articles to the Physics and Astronomy Classification System (PACS) and the Astronomical Subject Keywords would have been valuable additional sources of validation data and could perhaps be used in follow-up research.

Our results indicate that the solution sr of Boyack (2017) attains the overall best scores, albeit not uniformly so. On the special issues validation dataset the solutions of Glänzel & Thijs (2017) have the advantage. The sr solution achieves these results with clustering using direct citation links of a far larger dataset covering all sciences. Unlike the other solutions, this clustering uses data from outside of the *Astro* dataset itself. Boyack (2017) found that "the number of links to papers outside the astronomy dataset is roughly double that of the number of links within the set. Although some of the external signal is to other astronomy documents published in years not covered by the dataset, the external signal is significant and has a strong influence on the partitioning". Our results confirm that the additional signal improves the clustering quality.

The good results of both solutions of Glänzel & Thijs (2017) in the special issues data are most likely due to using bibliographic coupling. Interestingly, the solution adding NLP similarity does not obtain better results in the special issues data. In fact, it only clearly outperforms the bibliographic-coupling-only solution on the *also read* data.

Perhaps somewhat surprising is the performance of the solutions ok and ol, which were constructed with a highly sophisticated calculation of item similarities making use of a variety of features (Wang & Koopman, 2017). These solutions perform on par with others in two datasets, but are not competitive in the journal topics and special issues datasets. That is particularly curious, as this method explicitly uses journals as a feature and it is those two datasets in which item groupings are naturally constrained to journals which theoretically ought to give ok and ol some advantage.

Additionally, we can corroborate the observations of Velden, Boyack, et al. (2017) that the solutions of similar clustering approaches are much alike. Our results show that solutions based on similar clustering strategies also tend to obtain similar results in objective external validation. Namely, the sets (ol, ok), (en, eb), (c, u), and (mb, mc, md) generally perform very similarly across the four validation datasets with only a few exceptions. We also found that cluster quality is related to cluster content across several solutions for some topics but not for others.

By using several novel sources of validation data for document clustering we were able to show that clustering based on citation relations benefits strongly from including peripheral citation relations, that is, citations to and from the document set to be clustered, from publications not themselves part of the clustered document set. We also found some indication of a weakness of direct citation clusterings to properly group publications of one topic which were published simultaneously.

Acknowledgments

This research has made use of NASA's Astrophysics Data System. We would like to thank Anastasiia Tcypina and Nikolai Schmarbeck for help with data collection. We further thank Clarivate Analytics for granting permission to use the *Astro* dataset which is derived from the Web of Science database and for granting permission to make available the validation data which included Web of Science data. We also thank Michael J. Kurtz for his explanations of the ADS service.

Data availability

The original validation data is published at https://zenodo.org/record/4061694.

References

Ahlgren, P., Chen, Y., Colliander, C., & van Eck, N. J. (2020). Enhancing direct citations: A comparison of relatedness measures for community detection in a large set of PubMed publications. *Quantitative Science Studies*, *1*(2), 714–729. https://doi.org/10.1162/qss_a_00027

Bollen, J., Van de Sompel, H., Hagberg, A., Bettencourt, L., Chute, R., Rodriguez, M. A., & Balakireva, L. (2009). Clickstream data yields high-resolution maps of science. *PLoS One*, *4*(3), e4803. https://doi.org/10.1371/journal.pone.0004803

Boyack, K. W. (2017). Investigating the effect of global data on topic detection. *Scientometrics*, *111*(2), 999–1015. https://doi.org/10.1007/s11192-017-2297-y

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology*, *61*(12), 2389–2404. https://doi.org/10.1002/asi.21419

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., ... Börner, K. (2011). Clustering more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PloS ONE*, *6*(3). https://doi.org/10.1371/journal.pone.0018029

Glänzel, W., & Thijs, B. (2017). Using hybrid methods and 'core documents' for the representation of clusters and topics: The astronomy dataset. *Scientometrics*, *111*(2), 1071–1087. https://doi.org/10.1007/s11192-017-2301-6

Gläser, J., Glänzel, W., & Scharnhorst, A. (2017). Same data–different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, *111*(2), 981–998. https://doi.org/10.1007/s11192-017-2296-z

Halkidi, M., Vazirgiannis, M., & Hennig, C. (2015). Method-independent indices for cluster validation and estimating the number of clusters. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of cluster analysis* (pp. 616–639). Chapman & Hall/CRC.

Havemann, F., Gläser, J., & Heinz, M. (2017). Memetic search for overlapping topics based on a local evaluation of link communities. *Scientometrics*, *111*(2), 1089–1118. https://doi.org/10.1007/s11192-017-2302-5

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, *68*(4), 984–998. https://doi.org/10.1002/asi.23734

Koopman, R., Wang, S., & Scharnhorst, A. (2017). Contextualization of topics: Browsing through the universe of bibliographic information. *Scientometrics*, *111*(2), 1119–1139. https://doi.org/10.1007/s11192-017-2303-4

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., & Murray, S. S. (2005). Worldwide use and impact of the NASA Astrophysics Data System digital library. *Journal of the American Society for Information Science and Technology*, *56*(1), 36–45. https://doi.org/10.1002/asi.20095

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C. S., & Murray, S. S. (2002). Second-order bibliometric operators in the Astrophysics Data System. *Astronomical Data Analysis II*, *4847*, 238–245. https://doi.org/10.1117/12.460438

Kurtz, M. J., & Henneken, E. A. (2014). Finding and recommending scholarly articles. In B. Cronin & C. R. Sugimoto (Eds.), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact* (pp. 243–259). MIT Press.

Meila, M. (2015). Criteria for comparing clusterings. In C. Hennig, M. Meila, F. Murtagh, & R. Rocci (Eds.), *Handbook of cluster analysis* (pp. 640–657). Chapman & Hall/CRC.

Palchykov, V., Gemmetto, V., Boyarsky, A., & Garlaschelli, D. (2016). Ground truth? Concept-based communities versus the external classification of physics manuscripts. *EPJ Data Science*, *5*(1), 28. https://doi.org/10.1140/epjds/s13688-016-0090-4

Peel, L., Larremore, D. B., & Clauset, A. (2017). The ground truth about metadata and community detection in networks. *Science Advances*, *3*(5), e1602548. https://doi.org/10.1126/sciadv.1602548%20

Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, *9*(1), 102–117. https://doi.org/10.1016/j.joi.2014.11.010

Shu, F., Julien, C.-A., Zhang, L., Qiu, J., Zhang, J., & Larivière, V. (2019). Comparing journal and paper level classifications of science. *Journal of Informetrics*, *13*(1), 202–225. https://doi.org/10.1016/j.joi.2018.12.005

Sjögårde, P., & Ahlgren, P. (2018). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics. *Journal of Informetrics*, *12*(1), 133–152. https://doi.org/10.1016/j.joi.2017.12.006

Sjögårde, P., & Ahlgren, P. (2020). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of specialties. *Quantitative Science Studies*, *1*(1), 207–238. https://doi.org/10.1162/qss_a_00004

Šubelj, L., van Eck, N. J., & Waltman, L. (2016). Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PLOS ONE*, *11*(4), e0154404. https://doi.org/10.1371/journal.pone.0154404

van Eck, N. J., & Waltman, L. (2017). Citation-based clustering of publications using CitNetExplorer and VOSviewer. *Scientometrics*, *111*(2), 1053–1070. https://doi.org/10.1007/s11192-017-2300-7

Velden, T., Boyack, K. W., Gläser, J., Koopman, R., Scharnhorst, A., & Wang, S. (2017). Comparison of topic extraction approaches and their results. *Scientometrics*, *111*(2), 1169–1221. https://doi.org/10.1007/s11192-017-2306-1

Velden, T., Yan, S., & Lagoze, C. (2017). Mapping the cognitive structure of astrophysics by infomap clustering of the citation network and topic affinity analysis. *Scientometrics*, *111*(2), 1033–1051. https://doi.org/10.1007/s11192-017-2299-9

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, *11*, 2837–2854.

Waltman, L., Boyack, K. W., Colavizza, G., & van Eck, N. J. (2020). A principled methodology for comparing relatedness measures for clustering publications. *Quantitative Science Studies*, *1*(2), 691–713. https://doi.org/10.1162/qss_a_00035

Wang, S., & Koopman, R. (2017). Clustering articles based on semantic similarity. *Scientometrics*, *111*(2), 1017–1031. https://doi.org/10.1007/s11192-017-2303-4

Xu, S., Liu, J., Zhai, D., An, X., Wang, Z., & Pang, H. (2018). Overlapping thematic structures extraction with mixed-membership stochastic blockmodel. *Scientometrics*, *117*(1), 61–84. https://doi.org/10.1007/s11192-018-2841-4

Zhang, Y., Lu, J., Liu, F., Liu, Q., Porter, A., Chen, H., & Zhang, G. (2018). Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *Journal of Informetrics*, *12*(4), 1099–1117. https://doi.org/10.1016/j.joi.2018.09.004

Appendix

Table 9: Topical sections of four astronomy journals with occurrence counts

journal	sections
Astronomy & Astrophysics	Astronomical instrumentation (434); Astrophysical processes (647); Atomic, molecular, and nuclear data (260); Catalogs and data (156); Celestial mechanics and astrometry (185); Cosmology (including clusters of galaxies) (949); Extragalactic astronomy (2300); Galactic structure, stellar clusters, and populations (822); Instruments, observational techniques, and data processing (39); Interstellar and circumstellar matter (1740); Numerical methods and codes (53); Online catalogs and data (59); Planets and planetary systems (792); Stellar atmospheres (697); Stellar structure and evolution (2159); The Sun (1210)
Chinese Journal of Astronomy and Astrophysics	ASTROMETRY (2); ASTROMETRY AND CELESTIAL MECHANICS (11); ASTROPHYSICAL PROCESSES (8); CELESTIAL MECHANICS (2); COSMOLOGY (28); EXTRAGALACTIC ASTRONOMY (106); EXTRAGALACTIC COSMOLOGY (9); GALACTIC STRUCTURE AND DYNAMICS (4); HIGH ENERGY ASTROPHYSICS (57); HISTORY OF ASTRONOMY (4); INSTRUMENTS, OBSERVATIONAL TECHNIQUES AND DATA PROCESSING (29); INTERSTELLAR MEDIUM (7); JET SOURCES & GAMMA RAY BURSTS (14); NEUTRON STARS: ORIGIN AND EVOLUTION (13); PULSAR EMISSION THEORY (11); PULSAR OBSERVATIONS (15); PULSAR TIMING (10); STARS (143); STELLAR CLUSTERS (4); TECHNIQUES AND NEXT-GENERATION TELESCOPES (5); THE GALAXY (3); THE SOLAR SYSTEM (5); THE SUN (73); THE SUN AND SOLAR SYSTEM (2)
Kinematics and Physics of Celestial Bodies	Dynamics and Physics of Bodies of the Solar System (17); Earth: Rotation and Geodynamics (3); Extragalactic Astronomy (8); Extragalactic Astronomy and Cosmology (12); Instruments and Devices (4); Interstellar Medium and Nebulae (5); Physics of Stars and Interstellar Medium (12); Positional and Theoretical Astronomy (2); Problems of Astronomy (8); Solar Physics (19); Space Physics (5); Stars (2); Structure and Dynamics of the Galaxy (8)
Publications of the Astronomical Society of the Pacific	Astronomical Instrumentation (188); Astronomical Phenomena and Seeing (44); Astronomical Techniques (52); Atmospheric Phenomena and Seeing (19); Data Analysis and Techniques (98); Extrasolar Planets (37); Galaxies (43); Interstellar Medium and Nebulae (17); ISM (15); Quasars and Active Galactic Nuclei (11); Solar System (11); Star Clusters and Associations (41); Stars (259); Supernovae (20)

Table 10: Three best and worst performing clusters per solution (only values for clusters for which all four ratios could be calculated)

solution	cluster	cluster size	ratio NSF grants	ratio journal sections	ratio special issues	weighted ratio also read	rank in solution	average ratio
c	20	1,963	0.00	0.00	0.10	0.01	18	0.03
c	21	1,839	0.14	0.09	0.27	0.20	17	0.18
c	8	5,211	0.37	0.00	0.35	0.03	16	0.19
c	4	7,483	0.61	0.23	0.34	0.32	3	0.37
c	3	7,998	0.49	0.88	0.65	0.00	2	0.50
c	1	14,876	0.72	0.57	0.40	0.56	1	0.56
eb	13	9,538	0.17	0.00	0.46	0.00	13	0.16
eb	12	4,639	0.36	0.08	0.16	0.21	12	0.20
eb	7	5,755	0.35	0.16	0.06	0.25	11	0.21
eb	2	10,408	0.53	0.50	0.51	0.31	3	0.46
eb	1	10,666	0.59	0.43	0.55	0.35	2	0.48
eb	4	12,678	0.46	0.84	0.80	0.02	1	0.53
en	2	17,568	0.09	0.02	0.45	0.01	11	0.14
en	9	6,055	0.32	0.05	0.16	0.19	10	0.18
en	6	6,936	0.19	0.21	0.26	0.34	9	0.25
en	11	14,830	0.33	0.67	0.64	0.08	3	0.43
en	8	13,485	0.59	0.40	0.48	0.49	2	0.49
en	1	16,142	0.54	0.52	0.67	0.47	1	0.55
mb	87	1,152	0.01	0.00	0.02	0.01	107	0.01
mb	31	803	0.00	0.01	0.04	0.00	106	0.01
mb	41	873	0.04	0.00	0.02	0.00	105	0.01
mb	7	1,580	0.77	0.17	0.23	0.00	3	0.29
mb	2	1,206	0.57	0.07	0.32	0.34	2	0.32
mb	22	1,193	0.94	0.10	0.27	0.03	1	0.33
mc	95	915	0.00	0.00	0.03	0.00	109	0.01
mc	113	802	0.00	0.00	0.08	0.00	108	0.02
mc	94	423	0.01	0.03	0.01	0.04	107	0.02
mc	93	1,150	0.19	0.06	0.59	0.22	3	0.27
mc	12	614	0.50	0.05	0.20	0.33	2	0.27
mc	63	965	0.75	0.06	0.42	0.00	1	0.31
md	97	969	0.00	0.00	0.02	0.01	112	0.01
md	18	924	0.00	0.01	0.03	0.01	111	0.01
md	20	927	0.00	0.01	0.05	0.01	110	0.02
md	108	817	0.45	0.10	0.08	0.25	3	0.22
md	81	1,058	0.83	0.10	0.10	0.00	2	0.26
md	1	938	0.57	0.05	0.19	0.28	1	0.27

Table 10: Three best and worst performing clusters per solution (only values for clusters for which all four ratios could be calculated) *(continued)*

solution	cluster	cluster size	ratio NSF grants	ratio journal sections	ratio special issues	weighted ratio also read	rank in solution	average ratio
ok	19	1,627	0.05	0.00	0.02	0.00	27	0.02
ok	3	3,389	0.00	0.00	0.12	0.00	26	0.03
ok	5	3,874	0.00	0.00	0.16	0.02	25	0.05
ok	8	2,195	0.69	0.16	0.39	0.39	3	0.41
ok	2	5,449	0.57	0.28	0.47	0.37	2	0.42
ok	14	5,583	0.82	0.84	0.43	0.01	1	0.52
ol	29	639	0.00	0.00	0.01	0.17	29	0.05
ol	30	2,764	0.07	0.05	0.05	0.01	28	0.05
ol	26	6,095	0.20	0.02	0.33	0.01	27	0.14
ol	18	2,702	0.59	0.15	0.38	0.33	3	0.36
ol	12	6,893	0.45	0.33	0.30	0.39	2	0.37
ol	9	6,585	0.41	0.81	0.49	0.00	1	0.43
sr	104	407	0.00	0.02	0.06	0.01	19	0.02
sr	147	260	0.00	0.02	0.11	0.01	18	0.03
sr	238	457	0.07	0.02	0.04	0.14	17	0.07
sr	400	7,076	0.62	0.82	0.55	0.01	3	0.50
sr	126	19,988	0.75	0.66	0.35	0.61	2	0.59
sr	17	33,874	0.77	0.79	0.68	0.64	1	0.72
u	6	5,692	0.36	0.01	0.32	0.03	14	0.18
u	15	2,171	0.15	0.10	0.26	0.22	13	0.18
u	2	12,432	0.40	0.13	0.28	0.10	12	0.23
u	3	11,477	0.63	0.36	0.55	0.36	3	0.47
u	4	7,925	0.87	0.88	0.64	0.00	2	0.60
u	1	18,259	0.70	0.73	0.44	0.61	1	0.62