Mapping the Cognitive Structure of Astrophysics by Infomap Clustering of the Citation Network and Topic Affinity Analysis

Theresa Velden Shiyan Yan Carl Lagoze

February 9, 2017

Abstract

This is a preprint of an article accepted to be published in a special issue of Scientometrics: Gläser, J., Scharnhorst, A. and Glänzel, W. (eds), Same data – different results? Towards a comparative approach to the identification of thematic structures in science.

In this paper we use the information theoretic Infomap algorithm (Rosvall and Bergstrom, 2008) iteratively in order to cluster the direct citation network of the *Astro Data Set* (publications in 59 astrophysical journals between 2003-2010.) We obtain 22 clusters of documents from the giant component of the network that we interpret as constituting 'topics' in the field of astrophysics. Upon investigation of the content of the topics we find a grouping of topics by shared features of their 'journal signature', that is the journals that are most characteristic for a topic due to their popularity and distinctiveness. These groups of topics match sub disciplines within the field. We generate a cognitive map of the field using a topic affinity network that shows what topics are disproportionally well connected (by citations) to other topics. The topology of the topic affinity network highlights a high-level organization of the field by sub-discipline and observational distance of the research object from Earth.

1 Introduction

The mapping of social and cognitive structures in scientific research fields (Morris and Van der Veer Martens, 2008) is flourishing for a number of reasons. Increasingly, scholarly publications and their metadata are available from a variety of sources (digital libraries, institutional and disciplinary repositories, along with bibliographic abstracting services such as the long established Web of Science and more recently, Scopus). Complementing this is the emergence of sophisticated algorithms for the analysis of complex networks (Newman, 2003) and the wide availability of advanced user-friendly network analysis and visualization tools like pajek (Batagelj and Mrvar, 2003) and gephi (Bastian et al,

2009) for generic networks, or specialized tools for the analysis and visualization of scholarly networks, such as VOSviewer (Van Eck and Waltman, 2010), and CiteSpace (Chen, 2006).

However, many different approaches for community extraction and topic detection exist and suggest different answers what the most prominent groupings of authors or publications in a field are. The articles in this special issue set out to describe and compare various approaches to topic extraction from the scientific literature in order to evaluate the origin, extent, and implication of differences between methods (see the introduction to this special issue by Gläser et al (2017)). In this paper we use one specific method for topic detection and topic affinity analysis that we have previously introduced (Velden et al, 2010; Velden and Lagoze, 2013), and focus on describing the results we obtain if we apply this approach to the shared *Astro Data Set*.

Our approach has emerged from our research on communication and collaboration behaviors in scientific communities and their comparison across fields. We pursue a research program that takes a mixed method approach to studying field-specific practices and cultures of scientific communities, as described in Velden et al (2010); Velden and Lagoze (2013). In particular, we integrate ethnographic field studies with network analytic methods. This evolves a tradition of close-up analysis of scientific networks and communication practices started by Crane's work (Crane, 1972) on invisible colleges and taken up more recently by Zuccala (2006) and Cambrosio et al (2004). Our approach recognizes that scientific research specialties are a complex social and cognitive phenomenon (Ding, 2011). Sociologically, they can be characterized as collective knowledge production communities that emerge from the indirectly coordinated activity of autonomous actors (research groups) who aim to contribute to a shared knowledge base (Gläser, 2006; Velden and Lagoze, 2013). To map the social and cognitive structure in a field, we usually proceed in two steps: First we algorithmically extract major research topics in a research specialty from the direct citation network of articles. Then we generate an affinity network that shows what topics are disproportionally well connected through citations to other topics. In a second step, we would generate the co-author network, after author name disambiguation which is important to avoid distortions of the co-author network (Velden et al, 2011), and overlay the topic information on the group collaboration network that has been extracted from the co-author network of the research specialty (Velden et al, 2010). The resulting map would show how collaborative ties connect groups active in a particular topic area. In this paper, however, we only apply the first step of this procedure to provide results for a direct comparison with the topic extraction approaches of other groups in this special issue.

2 Method

Our approach to topic extraction and topic affinity analysis was first reported in Velden et al (2013). We here review the procedure with particular emphasis on explicating motivations and details that may be relevant when comparing our results with those obtained by the other approaches in this special issue.

Data
The data set used in this study includes publications published 2003-2010 in 59 astrophysical journals indexed by Web of Science. Limited to documents of type 'Article', 'Letter', and 'Proceedings Paper', the data set consists of the bibliographic data of 111,616 publications.

Network construction
We construct the direct citation network of publications in the *Astro Data Set*. This is based on the assumption that a citation link signals a direct topical relatedness between two publications. The giant component of this network includes 101, 831 publications (91.2% of the data set). The next smallest component has only 48 documents, indicating a large scatter of the documents that are not included in the giant component. Given such a large concentration of documents in the giant component we routinely restrict the topic extraction analysis to the giant component of the network. This is based on the assumption that anything not connected to the giant component resides at the fringes of this domain and is not relevant for the within domain community structures that we are usually interested in for our behavioral studies. For the analysis of the *Astro Data Set* this might imply that interdisciplinary topics, such as e.g. astro-biology or astro-chemistry, if they primarily cite journals outside the selected set of core journals in astrophysics, are hidden within the scatter of 9% of publications that are excluded from our analysis.

Various citation-based approaches have been used in the past to detect topics in research fields. These include modeling the data as either networks or similarity matrices that relate publications to one another[1] based on direct citation, bibliographic coupling, or co-citation links. At the time we developed our approach in 2008/2009, a study ny Shibata et al (2009) was published that compared the three citation based approaches. Based on an analysis of data from three fields, ranging in size from about 3, 000 to 30, 000 publications in the giant component of the citation networks, it concluded direct citation was most

[1]Technically a similarity matrix can always be interpreted as an adjacency matrix that specifies for each pair of nodes in a network the strength of their connection. There exists a practical difference however between cases where the data model operationalizes the relationship between entities by measuring some direct interaction, e.g. a citation from one document to another, versus cases where the relationship between entities is operationalized as a similarity, e.g. the similarity in how two documents are citing or being cited by all other documents in the data set. In the former case, the network will typically be sparse, whereas in the latter case the network is typically dense and commonly some threshold is applied to suppress weak links between nodes to make calculations on the network easier.

accurate for capturing research fronts. One of its observations, based on calculating time series of network modularity, is that the direct citation networks tend to be more sparse and locally more dense, whereas for bibliographic coupling and co-citation networks modularity is lower and the networks are more random, which encouraged us to go ahead with a direct citation network to model our data. In 2010, another comparative study based on a much larger data set (over 2 million articles published in bio-medicine over five years, 2004-2008) was published (Boyack and Klavans, 2010). It differed in several ways, e.g. by including citation links to external publications in the modeling, basing the data model on similarity matrices rather than networks, and using different measures for evaluating the accuracy of clusterings in the absence of a ground truth. It came to different conclusions. Interestingly, it showed that the textual coherence of topics extracted from the direct citation network were much smaller than for bibliographic coupling and for co-citation analysis. The authors indicated, however, that what represents the most accurate approach may depend on the area of investigation and the kind of topics one is interested in. Very recently, the same authors have published another comparative study of direct citation versus bibliographic coupling versus co-citation networks, using for the direct citation network a 15-year time window of data covering 48.5 million documents, including 24.6 million source documents indexed by the Scopus database (Klavans and Boyack, 2015). This time they conclude that direct citation outperforms the other two approaches if the objective is to create taxonomies of science that reflect long term manifestations of topics rather than capturing recent research fronts. The authors further state, that the shorter time window (5 years) in their previous study had disadvantaged the direct citation approach. While progress is being made, these divergent findings underline the still spotty evidence on the influence of data models and clustering algorithms on the results of topic extraction which has motivated us to join the collaborative effort that is documented in this special issue.

Clustering

We use the Infomap algorithm (Rosvall and Bergstrom, 2008) as part of our work flow to extract topics from direct citation networks. The Infomap algorithm attempts to capture the network structure with respect to the network dynamics and partitions the network by looking for a minimal description of a function that describes network structure in terms of flow, the so called map equation. This is conceptually different from modularity based clustering algorithms that infer cluster membership based on a model of the process by which the network was generated, and this difference can lead to different clustering results (Rosvall et al, 2010). We apply to the *Astro Data Set* the work flow that we developed in 2008/2009 with the then available version of the Infomap code. We have applied this work flow in the past to data sets representing other research fields (Velden et al, 2010; Velden and Lagoze, 2013; Velden et al, 2015). In the meantime, the Infomap clustering algorithm has seen further development, so if one was to implement this workflow today, it likely would look different. We will insert comments below to point to relevant new developments.

In our implementation of the work flow, we use the Infomap algorithm iteratively two times to generate clusters of clusters of publications based on the direct citation links between them[2]. In the following we will refer to these clusters of clusters of publications as 'clusters' for short, and for the purpose of this paper and the comparison of approaches in this special issue, we will interpret these clusters as **topics** extracted from the *Astro Data Set*.

We use the undirected version of the algorithm (infomap_undir), i.e. the network is treated as an undirected network. This is necessary, since the directed version of a citation network exposes a strong temporal ordering of links and if the directed version of Infomap is used on such a network it produces strongly fluctuating and hence unreliable results. This can be understood as follows. The temporal alignment of links in a citation network has two implications: there are many nodes that have no out-links (they represent publications, especially in the early years of the data set, that get cited but do not cite any older publications), and there are few loops (publications mutually citing each other). The Infomap algorithm models information flows on a network using a random walker. It settles on structures where the random walker gets rerouted for an extended time, however if there is a sparsity of loops such as in a time directed network, it reduces the detectability of those features. Also, whenever this random walker gets stuck in a directed network on a node that has no out-links it is randomly teleported to another node in the network. When random teleportation dominates over flow within clusters, the results become affected by large fluctuation [Martin Rosvall, private communication on February 11, 2009]. Using the undirected version of the code means that the network is transformed from a directed network into an undirected one[3]. Conceptually it implies that we consider the direction of a citation link to have no meaning for the degree of topical relatedness between two publications.

For the specific purpose that drove the design of our approach, namely to map scientific community structures in research specialties, we have found that to recursively iterate the clustering once is necessary to obtain sufficiently large entities (topics) for further visual inspection and analysis of their relationships. In the first round of clustering we obtain $1,996$ clusters. Only six of these clusters are greater than $1,000$ documents in size and together they cover only 11.4% of the publications in the giant component. Two thirds of the clusters comprise

[2]Today, the hierarchical generalization of the map equation introduced in (Rosvall and Bergstrom, 2011) makes it possible to cluster networks hierarchically with nested clusters in a principled way, see http://www.mapequation.org/code.html.

[3]Since we designed our workflow, alternatives to completely disregarding the directionality have become available, such as to limit the number of steps and perform unrecorded teleportation that does not influence the clustering (Lambiotte and Rosvall, 2012). According to Rosvall (private communication), using the flag –undirdir in the code provided at http://www.mapequation.org/code.html triggers a two-mode dynamics that assumes undirected links for calculating flows, but directed links when minimizing the code length. It has been used e.g. in (Mirshahvalad et al, 2012; West et al, 2016).

less than 50 publications each. Their small size and large number makes it unreasonable to assume they represent distinct sub-communities within the field. From the results of the first round of clustering we construct an undirected network of 1,996 nodes (clusters) with links weighted by simple counting of the citation links between clusters. Clustering this network of clusters we obtain 22 clusters, ranging in size from 18,259 down to 65 publications, see figure 1. We consider these in the following as the topics extracted from the network. As mentioned above, the second largest component of the full direct citation network (the largest component outside of the giant component) has only 48 documents, so it is smaller than even the smallest topic extracted by clustering the giant component. This supports our decision to focus our analysis exclusively on the giant component of the network.

Topic Affinity Network
To map the topical affinity between topics we construct a network using the set of documents representing topics as nodes and generating directed links between topics whenever their citation based 'affinity' is non-zero and positively valued. We calculate the affinity between topics relative to a null model that assumes a random distribution of citation links proportional to topic sizes. If the actual count of citation links from one topic to another exceeds the expected number of links, we insert a respective directed link in the affinity network. Hence, the existence of a link between topics in the affinity network indicates a surplus of connectivity between the two topics in question, whereas the absence of a link may either mean 'normal' (random) background connectivity or a negative affinity value ('antagonism'). Note that affinity as defined below is an asymmetric property, as we make use here of the directionality of citation links that we had to suppress, as explained above, for technical reasons in the clustering of the network. The advantage of considering the directionality of citation links in the topic affinity map is that when interpreting these maps one can appreciate asymmetries, that is cases when one topic builds on knowledge (cites) from the other, but the other topic does not reciprocate that interest. In other cases the exchange of knowledge is balanced and mutual.

The affinity of a topic (source s) for another topic (target t) is calculated as follows: We define the directed affinity AF_{ij} of topic s_i for topic t_j as the normalized difference between the actual count of citations from topic s_i to topic t_j and the expected count:

$$AF_{ij} = \frac{C_{ij} - E_{ij}}{\sqrt{E_{ij}}} \tag{1}$$

where the expected count E_{ij} is defined as follows:

$$E_{ij} := \frac{P_j}{\sum_{k \in A_{n-i}} P_k} \times \left(\sum\nolimits_{k \in A_{n-i}} C_{ik}\right)$$

A_{n-i} := all topics except topic s_i
P_j := number of publications in topic t_j

C_{ij} := number of citations from topic s_i into topic t_j

Topic Labeling
To support the interpretation of the topic affinity network, we use a semi-automatic approach to labeling topics. To this end, we analyze the frequency of journals that the publications in each topic are published in. Using a measure that combines the popularity of a journal within a topic (its share among the publications in the topic) with its idiosyncrasy (the share the topic has in occurrences of the journal across all topics), we produce a ranked list of the 15 most distinctive journals in each topic. In the past (Velden and Lagoze, 2013) we have used those ranked journal lists to derive topic labels from journal titles and found that they oftentimes reflect sub-disciplinary orientation of topics.

The distinctiveness score d of journals represented in a cluster is calculated as follows:

$$d = 10 * log(1 + (share * idiosyncrasy/100)) \quad (2)$$

where:

$$share := f/s$$

$$idiosyncrasy := f/F$$

s := number of publications in cluster
f := frequency of journal within cluster
F :=frequency of journal across all clusters

In addition, OCLC provided us and the other groups as part of our collaborative effort to compare approaches to the extraction of topics from the *Astro Data Set*, with two further methods for inspecting the content of topics: First, 'Little Ariadne' (Koopman et al, 2017), an interactive web interface for investigating the *Astro Data Set*. It extracts terms and subjects from the publications belonging to a cluster. Second, a list of natural language terms for each topic, extracted from titles and abstracts using a method further described in Koopman and Wang (2017). We list those terms in table **??**.

3 Results

The topic extraction from the giant component of the direct citation network results in 22 clusters of publications (or topics). The distribution of cluster sizes is given in figure 1.

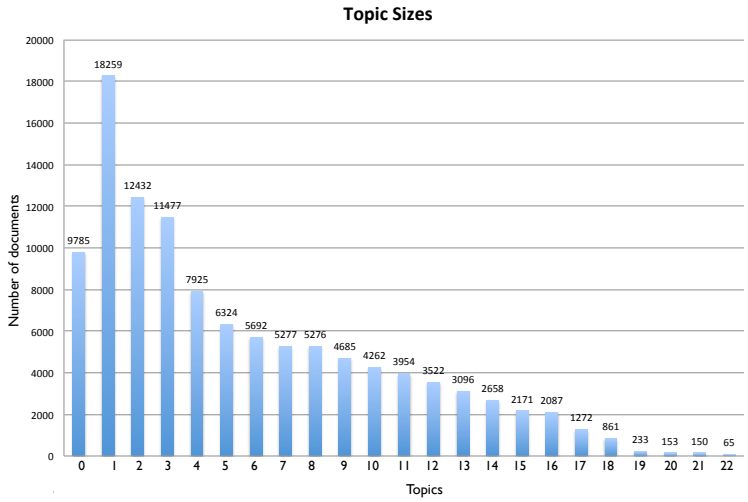

Figure 1: Sizes of the 22 topics (sets of documents) that constitute the giant component of the direct citation network. Cluster 0 shows the number of documents not included in the giant

3.1 Stability of clustering

Many clustering algorithms that are based on function optimization have a stochastic component (Möller, 2005). The Infomap algorithm that we use implements a fast stochastic and recursive search algorithm in order to minimize the so called map equation (Rosvall and Bergstrom, 2009). Due to the randomness of results introduced by this stochastic aspect of the algorithm, it is generally recommended to run it repeated times to increase the level of stability, and we used the -N flag to run it 10 times. We decided to probe the level of variation that remained, even when using this flag[4]. Specifically, we repeated the original clustering run (labeled UMSI0) twice on a different machine than the original run: Once with the same parameter value p for the random seed like in the original run ($p = 3451234$), producing a solution UMSI6 with 25 clusters, and once with a different value for the seed ($p = 123$) producing solution UMSI7 with 21 clusters. This shows that using the same seed but on a different machine, as well as using a different seed on the same machine produce solutions with different numbers of clusters. The heatmaps in figure 2 show overlap between solutions, where overlap is defined as an asymmetric measure of how much a cluster from one solution is covered by a cluster from the other solution. A light colored cell indicates that a high percentage of documents in the respective cluster of the source solution on the y-axis is included in the corresponding

[4] A more systematic and comprehensive method for determining variation in the form of a significance analysis is described in Rosvall and Bergstrom (2010).

cluster of the target solution on the x-axis. The split of a cluster of the original solution into two clusters is indicated in such a heatmap e.g. by a pattern like the one found in the top heatmap, comparing UMSI6 with UMSI0: The light color in the cells $s25/t3$ and $s3/t3$ indicate that cluster t3 has been split into a large cluster s3 and a much smaller cluster s25. Note that the clusters are numbered and ranked by cluster size within the solution: the lower the number the larger the relative size of the cluster within the solution.

The first two heatmaps in figure 2 show that there is a high percentage of overlap between clusters when compared to the original solution UMSI0, indicating great similarity of the solutions. Specifically, 23 clusters of the 25-cluster solution UMSI6 overlap by at least 90% with clusters in the original solution UMSI0. In UMSI7, 17 clusters overlap by at least 90% with clusters in the original solution. The main differences that can be derived from the heatmap is that in UMSI6 three clusters of the original UMSI0 solution have been split (UMSI0: t3, t6, and t13) resulting in 25 instead of 22 clusters. In UMSI7 two clusters of the original solution have been split (UMSI0: t2, and t6), and three new clusters were generated by merging clusters or subsets thereof (UMSI0: t10 and t16, t3 and t15, t17 and t18), resulting in 21 clusters, instead of 22.

For comparison, the third heatmap, at the bottom of figure 2 shows the overlap of a clustering result provided by Van Eck and Waltman (2017) (in this issue) with our original solution. Van Eck's solution (labeled CWTS-C5) was generated by using a different clustering algorithm on the same data using the same data model (direct citation network). There are significant overlaps between the two solutions UMSI0 and CWTS-C5, as 12 clusters from CWTS-C5 overlap by more than 90% with clusters in UMSI0 including the four largest clusters in CWTS-C5. However the heatmap shows more splits and merges, and in particular a stronger scattering away from the main diagonal. This indicates that a number of the clusters in CWTS-C5 that have strong overlap with clusters in the original solution do no longer coincide in size rank, a feature that was strongly present in the first two heatmaps that show variations due to the stochastic character of the Infomap code. The visualization highlights in particular, that the fifth biggest cluster from the original solution UMSI0 has been split into two similarly sized clusters, clusters 12 and 14 from the CWTS-C5 solution.

3.2 Content of Topics

Tables 1 and 2 list for each topic the 5 most characteristic journals, ranked by distinctiveness score (see equation 2). Comparing those lists between topics, we find that groups of topics share striking similarities in their journal data, namely the titles, frequency, share, and idiosyncrasy of the most distinctive journals in a topic. In the following we refer to this data as the 'journal signature' of a topic. A high level grouping of topics is suggested by their journal signatures. We derived labels for these groups of topics from journal titles in combination with inspection of topic content using the meta data provided by 'Little Ari-

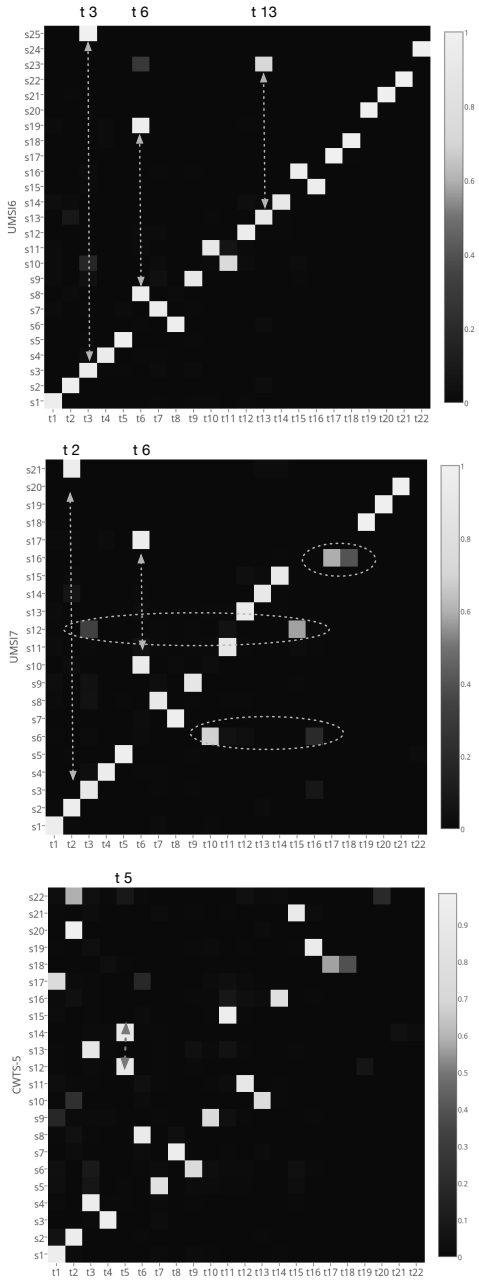

Figure 2: Each heatmap indicates the proportion of overlap of clusters in one solution on the y-axis (from top to bottom: UMSI6, UMSI7, CWTS-C5) with the original solution 'target solution' on the x-axis (UMSI0). Possible values range from 0 (no overlap) to 1.0 (complete inclusion). The clusters are labeled by size, so clusters sizes decrease as one moves right on the x-axis and up on the y-axis. Dotted lines indicate splits and merges of clusters in the target solution discussed in the text

adne' (Koopman et al, 2017) and used them to organize the topics in tables 1 and 2.

It took some time to learn how to interpret the journal signatures and to decide what topical groupings to distinguish and what topics to assign to each group. It meant to eye the data given in tables 1 and 2 about the most popular and distinct journals in each topic, and look for commonalities between topics and striking differences. When a suggestive pattern was found, we verified the relatedness of the topics that shared this pattern by inspection of topic contents in 'Little Ariadne.' In retrospect, taking all the available information into account, the distinguishing characteristics of the journal signatures of topics that belong to a group seem more obvious and the assignment of topics to a group is straightforward in most cases.

An initially subtle step was to decide to separate group 1 (Gravitational Physics, Cosmology) and group 2 (Astroparticle Physics), since Physical Review D is an important journal in both domains that ranks for each topic among the top two journals. However, the Astroparticle Physics domain is distinct in the very high concentration of publications in this one journal - both topics listed as Astroparticle physics have a large majority of publications concentrated in Physical Review D (72 and 98%), whereas within the gravitational physics domain the journal has at most a 45% share. Inspection of topic contents with Little Ariadne (and similarly the terms displayed in table **??**) support making this distinction as mainly high energy physics terms and subjects are provided for those two topics whereas terms indicative of gravitational physics are missing.

Group 3 (Astrophysics) is a large domain, with a handful of large journals dominating the output in this field. For each topic in this group, the following journal titles together subsume two-thirds or more of publications: Astrophysical Journal, Monthly Notices of the Royal Astronomical Society, Astronomy & Astrophysics, Astronomical Journal.

Group 4 (Solar Physics) incorporates a single, rather large topic (the 4th biggest cluster with $6,324$ documents). It has with 1,248 documents a 93% share of all the publications published in the journal 'Solar Physics'. This journal signature suggests a very specialized and clearly delineated topic. Whereas a larger portion of documents (3,774) in this topic is published in the large journals Astrophysical Journal and Astronomy & Astrophysics that are prominent also in topics of group 3, the important distinction here is that Solar Physics tops the ranked journal list in terms of distinctiveness score d, and that the other two journals are so broad in scope that it is very plausible that work in solar physics would get published in these more general journals.

Group 5 (Planetary Science) primarily accounts for the fifth largest topic, a cluster of documents where all 5 top journal are highly specialized and almost

Table 1: Grouping of topics by journal signature. (Part 1)

topic	source title	f	share	idios.	d
	Group 1: Gravitational Physics & Cosmology				
2	Physical Review D	5616	45.2	30.9	0.014
	Journal of Cosmology And Astroparticle Physics	1416	11.4	74.1	0.008
	Classical And Quantum Gravity	1533	12.3	42.7	0.005
	International Journal of Modern Physics D	655	5.3	54.4	0.003
	General Relativity And Gravitation	543	4.4	58.9	0.003
13	Classical And Quantum Gravity	875	28.3	24.4	0.007
	Physical Review D	1773	57.3	9.7	0.006
	General Relativity And Gravitation	229	7.4	24.8	0.002
	International Journal of Modern Physics D	146	4.7	12.1	0.001
	Space Science Reviews	2	0.1	0.6	0.000
14	Classical And Quantum Gravity	1024	38.5	28.5	0.011
	Physical Review D	1022	38.4	5.6	0.002
	General Relativity And Gravitation	95	3.6	10.3	0.000
	International Journal of Modern Physics D	51	1.9	4.2	0.000
	Astrophysical Journal	165	6.2	0.8	0.000
	Group 2: Astroparticle Physics				
6	Physical Review D	4101	72.0	22.5	0.016
	Astroparticle Physics	430	7.6	61.8	0.005
	Journal of Cosmology And Astroparticle Physics	353	6.2	18.5	0.001
	Astrophysical Journal	241	4.2	1.2	0.000
	New Astronomy Reviews	45	0.8	5.6	0.000
8	Physical Review D	5208	98.7	28.6	0.028
	Nuovo Cimento C-Geophysics And Space Physics	3	0.1	2.2	0.000
	New Astronomy Reviews	1	0.0	0.1	0.000
	Journal of Cosmology And Astroparticle Physics	5	0.1	0.3	0.000
	International Journal of Modern Physics D	31	0.6	2.6	0.000
	Group 3: Astrophysics				
1	Monthly Notices of The Royal Astronomical Society	4415	24.2	38.3	0.009
	Astrophysical Journal	5565	30.5	28.5	0.009
	Astronomy & Astrophysics	3148	17.2	21.7	0.004
	Astronomical Journal	1098	6.0	32.4	0.002
	Astrophysical Journal Supplement Series	401	2.2	40.2	0.001
3	Astronomy & Astrophysics	3122	27.2	21.5	0.006
	Astrophysical Journal	2773	24.2	14.2	0.003
	Monthly Notices of The Royal Astronomical Society	1461	12.7	12.7	0.002
	Astronomical Journal	732	6.4	21.6	0.001
	Publications of The Astronomical Society of The Pacific	364	3.2	40.3	0.001

Table 2: Grouping of topics by journal signature. (Part 2)

topic	source title	f	share	idios.	d
	Group 3: Astrophysics (Cont'd)				
7	Astrophysical Journal	1856	35.2	9.5	0.003
	Astronomy & Astrophysics	1359	25.8	9.4	0.002
	Monthly Notices of The Royal Astronomical Society	686	13.0	6.0	0.001
	Astrophysics And Space Science	176	3.3	9.0	0.000
	Astrophysical Journal Supplement Series	131	2.5	13.1	0.000
9	Astronomical Journal	571	12.2	16.9	0.002
	Astronomy & Astrophysics	1051	22.4	7.2	0.002
	Monthly Notices of The Royal Astronomical Society	909	19.4	7.9	0.002
	Astrophysical Journal	1073	22.9	5.5	0.001
	Publications of The Astronomical Society of Australia	86	1.8	27.8	0.001
10	Astrophysical Journal	1332	31.3	6.8	0.002
	Astronomy & Astrophysics	897	21.0	6.2	0.001
	Monthly Notices of The Royal Astronomical Society	783	18.4	6.8	0.001
	Publications of The Astronomical Society of Japan	217	5.1	19.5	0.001
	Chinese Journal of Astronomy And Astrophysics	82	1.9	12.5	0.000
11	Astrophysical Journal	1459	36.9	7.5	0.003
	Nuovo Cimento C-Geophysics And Space Physics	105	2.7	75.5	0.002
	Monthly Notices of The Royal Astronomical Society	596	15.1	5.2	0.001
	Astronomy & Astrophysics	589	14.9	4.1	0.001
	Astrophysical Journal Letters	162	4.1	7.4	0.000
12	Astrophysical Journal	1097	31.1	5.6	0.002
	Young Neutron Stars And Their Environments	68	1.9	86.1	0.002
	Monthly Notices of The Royal Astronomical Society	559	15.9	4.9	0.001
	Astronomy & Astrophysics	585	16.6	4.0	0.001
	Chinese Journal of Astronomy And Astrophysics	111	3.2	16.9	0.001
15	Planetary Nebulae: Their Evolution And Role	44	2.0	83.0	0.002
	Astronomy & Astrophysics	587	27.0	4.0	0.001
	Astrophysical Journal	536	24.7	2.7	0.001
	Revista Mexicana De Astronomia Y Astrofisica	43	2.0	20.7	0.000
	Monthly Notices of The Royal Astronomical Society	330	15.2	2.9	0.000
16	Monthly Notices of The Royal Astronomical Society	416	19.9	3.6	0.001
	Astronomy & Astrophysics	461	22.1	3.2	0.001
	Baltic Astronomy	61	2.9	16.6	0.001
	Astrophysical Journal	462	22.1	2.4	0.001
	Astronomical Journal	155	7.4	4.6	0.000

Table 3: Grouping of topics by journal signature. (Part 3)

topic	source title	f	share	idios.	d
	Group 4: Solar Physics				
4	Solar Physics	1248	15.7	93.9	0.015
	Astrophysical Journal	2165	27.3	11.1	0.003
	Astronomy & Astrophysics	1609	20.3	11.1	0.002
	Advances In Space Research	372	4.7	25.5	0.001
	Geophysical And Astrophysical Fluid Dynamics	77	1.0	68.1	0.001
	Group 5: Planetary Science				
5	Icarus	2102	33.2	92.2	0.031
	Planetary And Space Science	850	13.4	79.1	0.011
	Astrobiology	258	4.1	79.9	0.003
	Earth Moon And Planets	257	4.1	74.7	0.003
	Solar System Research	167	2.6	64.0	0.002
19	Celestial Mechanics & Dynamical Astronomy	95	40.8	24.2	0.010
	Jbis-Journal of The British Interplanetary Society	19	8.2	24.1	0.002
	Astrophysics And Space Science	35	15.0	1.8	0.000
	Earth Moon And Planets	10	4.3	2.9	0.000
	Astronomy Reports	14	6.0	1.9	0.000
20	Celestial Mechanics & Dynamical Astronomy	43	28.1	11.0	0.003
	Astronomy Reports	13	8.5	1.8	0.000
	Astronomy & Astrophysics	52	34.0	0.4	0.000
	Solar System Research	3	2.0	1.1	0.000
	New Astronomy	2	1.3	0.4	0.000
	Group 6: Space Science				
17	Annales Geophysicae	856	67.3	51.5	0.035
	Advances In Space Research	148	11.6	10.1	0.001
	Plasma Processes In The Near-Earth Space	13	1.0	68.4	0.001
	Solar Wind-Magnetosphere-Ionosphere Dynamics	10	0.8	76.9	0.001
	Planetary And Space Science	90	7.1	8.4	0.001
18	Annales Geophysicae	436	50.6	26.2	0.013
	Advances In Space Research	213	24.7	14.6	0.004
	Iri: Quantifying Ionospheric Variability	30	3.5	96.8	0.003
	Description of The Low Latitude And Equatorial Ionosph.	23	2.7	100.0	0.003
	Advances In Specifying Plasma Temperatures	23	2.7	92.0	0.003
21	Space Life Sciences: Missions To Mars	17	11.3	81.0	0.009
	Space Life Sciences: Closed Artificial Ecosystems	10	6.7	100.0	0.007
	Space Life Sciences: Gravity-Related Effects On Plants	9	6.0	100.0	0.006
	Space Life Sciences: Closed Ecological Systems	8	5.3	100.0	0.005
	Space Life Sciences: Life Support Systems	6	4.0	85.7	0.003
22	Space Life Sciences: Ground-Based Iron-Ion Biology	7	10.8	58.3	0.006
	Space Life Sciences: Radiation Risk Assessment	8	12.3	40.0	0.005
	Space Life Sciences: Flight Measurements	5	7.7	38.5	0.003
	Space Life Sciences: Aircraft And Space Radiation Env.	3	4.6	42.9	0.002
	Space Life Sciences: Structure And Dynamics of The Global	2	3.1	28.6	0.001

exclusive to this cluster. Two additional, very small topics were subsumed into this group, although the signal based on journal signature is not strong. Topics 19 and 20 are really focused on orbits (of planets, asteroids or spacecraft within the planetary system) - so by topic an assignment to planetary science seems a plausible decision, although creating a distinct group on 'Celestial Mechanics' or subsuming to group 6 would seem alternative options. The two topics were grouped into planetary science since the top ranked journal in both clusters (Celestial Mechanics and Dynamical Astronomy) has its largest share in topic 5 where 43% of publications in the journal are included, a fact one only sees if one considers an extended rank list, as this journal is at rank 6 for cluster 5 and hence cut off from the top 5 lists.

Finally, group 6 (Space Science) is composed of two subgroups of clusters. The two largest ones have a clear journal signature, with more than three quarters of articles in the clusters published in two journals: Annales Geophysicae and Advances in Space Research. The other two very small clusters are included in this group due to the titles of the 5 top ranked journals, all starting with 'Space Life Sciences'. These two small clusters are peculiar in that the top 5 clusters only unite less than 40% of all articles in the cluster. This looks like an artifact of the publication format and how it was indexed in the source field of the bibliographic records: the journal titles suggest a these are closely related series that could be considered a single journal. The 6th ranked journal for both clusters is Advances in Space Research. In both cases it contains more documents than the top five ranked journals (for cluster 21, 64 documents in Advances in Space Research versus 50 documents in total in the top five journals; and for cluster 22, 28 documents in Advances in Space Research versus a total of 25 documents in the top five journals) providing support for grouping the two clusters along with the other two into 'Space Research'. If share is calculated including the 6th ranked journal it jumps to more than 75%, corroborating that the low concentration in the top 5 journals is likely an artifact of how the source field was indexed.

When we combine the grouping of clusters suggested by journal signatures with a description of cluster content by the extracted natural language terms listed in table **??** we obtain a good sense of the granularity and orientation of topics extracted by our approach from the *Astro Data Set*. We will use this information in the next section, when discussing the cognitive topology of the field as depicted by the topic affinity network.

3.3 Topic Affinity Network

The topic affinity network of the 22 document clusters in solution UMSI0 that were extracted from the giant component of the direct citation network is shown in figure 3. Nodes represent topics and their coloring indicates membership in one of the topical groupings from tables 1, 2, and 3. Links between nodes indicate affinity between topics in terms of surplus of citations. Our first obser-

Table 4: Topic content descriptions (Part 1).

topic	word-based labels
Group 1: Gravitational Physics & Cosmology	
2	inflation, dark energy, microwave background, cosmic microwave, cosmological, universe, scalar field, gravity, cmb, background cmb
13	ads, black holes, horizon, spacetimes, hole solutions, quasinormal modes, supergravity, dimensional, hawking radiation, anti
14	gravitational wave, lisa, inspiral, binary black, wave detectors, ligo, laser interferometer, numerical relativity, post newtonian, waveforms
Group 2: Astroparticle Physics	
6	standard model, neutrino, higgs, lhc, minimal supersymmetric, lepton, supersymmetric standard, gev, muon, top quark
8	qcd, quark, meson, lattice, decays, chiral, pi pi, gluon, j psi, pion
Group 3: Astrophysics	
1	galaxies, redshift, active galactic, agn, star formation, galactic nuclei, quasar, sample, gas, galaxy clusters
3	star, planets, hd, main sequence, brown dwarfs, radial velocity, planet formation, transit, type stars, extrasolar planets
7	molecular cloud, protostellar, cloud, interstellar, star forming, young stellar, molecules, forming region, massive star, c 13
9	globular clusters, fe h, metal poor, giant branch, red giant, metallicity, stars, milky way, dwarf spheroidal, galactic globular
10	ray binary, x ray, neutron star, black hole, hard state, rossi x, timing explorer, ray timing, ultraluminous x, rxte
11	grb, ray bursts, gamma ray, afterglow, bursts grbs, sn, explosion, type ia, swift, supernova
12	pulsar, supernova remnant, psr, snr, neutron stars, wind nebula, anomalous x, radio pulsars, magnetar, remnant snr
15	planetary nebulae, pne, post agb, asymptotic giant, mira, central star, nebulae pne, agb stars, pn, symbiotic
16	white dwarf, nova, cataclysmic variable, dwarf nova, subdwarf b, wd, sdb stars, orbital period, sdb, superhumps
Group 4: Solar Physics	
4	solar, coronal mass, active region, cme, flare, magnetic field, mass ejections, sunspot, quiet sun, chromosphere

Table 5: Topic content descriptions (Part 2).

topic	word-based labels
Group 5: Planetary Science	
5	mars, comet, asteroid, titan, saturn, cassini, albedo, icarus, jupiter, ice
19	body problem, restricted three, periodic orbits, three body, photogravitational, equilibrium points, sail, collinear, sitnikov, thrust
20	nutation, iau, celestial reference, p03, iers, cip, capitaine, celestial mechanics, chandler, mathews
Group 6: Space Science	
17	auroral, substorm, magnetopause, ionospheric, cluster spacecraft, plasma sheet, magnetosheath, field aligned, superdarn, dayside
18	ionospheric, iri, degrees n, tec, electron content, ionosonde, summer, total electron, fof2, winter
21	life support, plant, wheat, food, bioregenerative, crops, waste, biomass, cultivation, ecological
22	dose, hzetrn, dosimetry, radiobiological, straggling, phits, station iss, polyethylene, fluka, protection

vation is that the organization of the network (i.e. what topics are connected to each other) reflects the topical grouping of topics. The affinity network of topics belonging to groups 1 (Gravitational Physics, Cosmology), 2 (Astroparticle Physics) and 3 (Astrophysics) is rather dense showing strong affinity between topics. The two topics of group 2 (Astroparticle Physics) seem 'embedded' in the Astrophysics group of topics although cluster 6 also shows some weak reciprocal links with topic 2 in the Gravitational Physics, Cosmology group of topics. Attached to this densely connected network area is a bifurcated tail of topics in group 4 (Solar Physics) and 5 (Planetary Science). A noticeable feature is that the largest topics in those two groups, topics 4 and 5, do not have reciprocal affinity links with the topics in the other groups, indicating that they build on (cite) documents in Astrophysics and Astroparticle Physics but do not provide a significant knowledge base for those topics in return. Topics in group 6 (Space Research) are only weakly attached to topics from other groups, and hence reside at the periphery of the network.

When we combine the information about the grouping of topics into topical groups given in tables 1, 2, and 3 with the information about the content of individual topics given by the natural language terms in tables 3.2 and 3.2, we can make the following observations about the cognitive structure of the field as reflected by the topic affinity map. At a high level of organization, groups of topics connect in a roughly elongated structure. Moving along this structure from left to right corresponds to moving from research objects at a larger

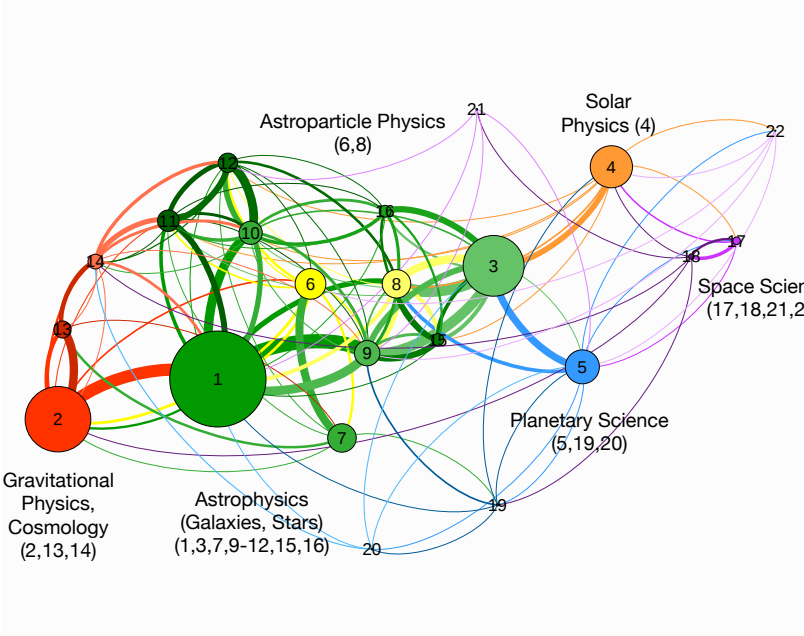

Figure 3: Topic affinity network. Node size indicates number of documents, and link strength relative preference given by publications in one topic to cite publications in another. Links are directed: they are colored by their source node and curve clockwise away from it. The colors of nodes - visible in the online version - indicate membership in a topical group as follows: **red** (Gravitational Physics, Cosmology), **yellow** (Astroparticle Physics), **green** (Astrophysics (stars, galaxies)), **orange** (Solar Physics), **blue** (Planetary science), **purple** (Space science). This network visualization was produced with gephi using the *Force Atlas 2* algorithm, one of the few network layout algorithms that considers edge weights in directed networks

scale of space-time to a smaller scale: It starts on the left with gravitational physics - the quest for understanding the workings of gravitational forces in the universe, and cosmology - the study of the origin of the universe. Next is astrophysics - the quest for developing a theoretical understanding of physical and chemical properties of celestial bodies. Within the astrophysics group of topics, the characteristic natural language terms allow us to distinguish that we first encounter topic 1 which is dedicated to galaxies, then topic 9 dedicated to globular clusters of stars in the galactic halo of galaxies, before moving on to topic 3 which is dedicated to smaller constituents, namely stars and the search for planetary systems surrounding them. We then narrow in on our local neighborhood, the sun and the solar system: topic 4 on solar physics - the dedicated study of the local star our solar system, and topic 5 on planetary science - the quest for understanding the composition, dynamics and history of planets in our solar system. Eventually, topics 17 and 18 on space science bring us to 'human' dimensions of space, regions that are accessible by spacecrafts from earth.

The topic affinity network further underlines that there are strong connections between some areas in astrophysics and gravitational physics. For example, a sub-group of smaller topics in *Group 3: Astrophysics*, namely topic 10 (neutron star and black hole binaries), topic 11 (supernovae), and topic 12 (pulsars and supernova remnants), are linked with topic 14 (gravitational waves) in *Group 1: Gravitational Physics & Cosmology*. This makes sense, because topics 10,11, and 12 focus on objects or systems that are of relevance to gravitational physicists as sources for gravitational waves. Other topics in astrophysics such as topic 3 (stars and extrasolar planets), 7 (star formation), 9 (globular clusters) and 15 (stellar evolution) do not show these links, as would be expected since they are focused on less gravitationally intense phenomena. Also, the cognitive links between gravitational physics and planetary science are weak.

4 Discussion

Interpreting the document clusters we extracted from the direct citation network as topics and using the affinity network approach to highlight their relationships, we arrive at a mapping of cognitive structures in the field that makes intuitively sense as it reveals an ordering of topics by the scale or distance in space-time of the research object being studied. Further, the fact that the journal signatures of topics suggest such a clear grouping of topics that aligns with sub-disciplines is an interesting and potentially useful finding. Given that we model the data as a direct citation network, the dominance of specific journals within a document cluster or a group of clusters might be accentuated by journal self-citation bias. However to what extent this bias is an indication of journal specialization (Rousseau, 1999) (and hence topical relatedness) or other factors such as attempts to manipulate journal impact factors (Dong et al, 2005), is hard to say. We would suggest that the dominant signal from citations is where researchers found knowledge they deemed relevant and useful for their research and this

would imply that the affinity network reveals a disciplinary organization of the shared knowledge base in the discipline with distance of the research object in space-time as a primary ordering principle.

Applying our approach for the extraction of topical structures to the *Astro Data Set* resulted in a coarse partition of the field into document clusters. Based on our own insights into the field of gravitational physics[5], we suggest that the document clusters we obtain fall short of distinguishing topics that researchers in the field would see as distinct, e.g. when discussing significant changes of focus in their research career. Our topic extraction distinguished only three topics within the 'Gravitational Physics and Cosmology' group of topics: topic 2 (cosmology), topic 13 (black holes), and topic 14 (gravitational waves). By contrast, one of the major review journals in the field of gravitational physics, *Living Reviews in Relativity*, is organized into nine subject categories: Experimental Foundations of Gravitation, Gravitational Waves, History of Relativity, Mathematical Relativity, Numerical Relativity, Physical Cosmology, Quantum General Relativity, Relativity in Astrophysics, String Theory and Gravitation. To leave out the second round of clustering in order to achieve greater granularity would increase resolution drastically as it results in almost 2,000 document clusters. However, as mentioned in the methods section, two thirds of these clusters are smaller than 50 documents in size. Hence, leaving out the second clustering step does not seem a viable option.

This suggests, that - depending on the desired level of resolution which will depend on the purpose of the topic extraction - one would have to explore how to achieve a more fine-grained clustering. In this context it would be very interesting to apply the generalized hierarchical version of Infomap described in Rosvall and Bergstrom (2011) to see whether it offers a plausible, intermediate result, more detailed than the partition we received with our iterative approach, but less dispersed than the partition we received from the one time application of Infomap that we started with. A systematic exploration of solutions produced with the same method but at various resolutions has been outside the scope of this work. Further, the question discussed in the introduction of this special issue by Gläser et al (2017) of what qualifies as a topic, remains. Without considering the specific purpose of a topic extraction exercise we lack an important criterion to assess the appropriateness of an approach and the results it produces.

5 Conclusions

The topology of the affinity network suggests cognitive links between the document clusters extracted by our method from the *Astro Data Set*. The organization of the network matches intuitively a cognitive map of the discipline

[5]One of the authors was trained in gravitational physics and worked several years as managing editor of the scientific review journal *Living Reviews in Relativity*.

that is based on distinctions in the scale or space-time context of the research object being studied, from large (universe) to small (our solar system). The interesting question in the context of the comparison of topic extraction approaches is whether the topics extracted by other approaches will aggregate documents into topics in significantly different ways than the topics we have extracted. Will the citation based affinity network, if produced from the document clusters extracted by the other approaches, merely reproduce the high level cognitive structure we see in our network? Or will we see some distinctive topological features that will be instructive to study in order to understand the nature of differences between the results of our topic extraction approaches? These questions will be addressed as part of the analysis in the comparison paper (Velden et al, 2017) in this special issue.

Acknowledgements

We gratefully acknowledge funding from SMA 1258891 EAGER: Collaborative Research: Scientific Collaboration in Time, as well as a travel grant by the intergovernmental framework for European Cooperation in Science and Technology (COST, Action: TD1210). We further thank Martin Rosvall for comments on pertinent new developments of the Infomap algorithm.

References

Bastian M, Heymann S, Jacomy M, et al (2009) Gephi: an open source software for exploring and manipulating networks. ICWSM 8:361–362

Batagelj V, Mrvar A (2003) Analysis and visualization of large networks. In: Graph Drawing Software, Springer, Berlin, pp 77–103

Boyack KW, 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

Cambrosio A, Keating P, Mogoutov A (2004) Mapping collaborative work and innovation in biomedicine a computer-assisted analysis of antibody reagent workshops. Social Studies of Science 34(3):325–364

Chen C (2006) Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. Journal of the American Society for Information Science and Technology 57(3):359–377

Crane D (1972) Invisible Colleges - Diffusion of Knowledge in Scientific Communities. The University of Chicago Press

Ding Y (2011) Community detection: Topological vs. topical. Journal of Informetrics 5(4):498–514

Dong P, Loh M, Mondry A (2005) The" impact factor" revisited. Biomedical digital libraries 2(7):1–8

Gläser J (2006) Wissenschaftliche Produktionsgemeinschaften - die soziale Ordnung der Forschung, Campus Forschung, vol 906. Campus Verlag, Frankfurt / New York

Gläser J, Glänzel W, Scharnhorst A (2017) Towards a comparative approach to the identification of thematic structures in science. In Gläser, J., Scharnhorst, A. & Glänzel, W. (eds), Same data – different results? Towards a comparative approach to the identification of thematic structures in science, Special Issue of Scientometrics X(Y):XYZ, doi:10.1007/s00000-000-0000-0

Klavans R, Boyack KW (2015) Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? ArXiv e-prints 1511.05078

Koopman R, Wang S (2017) Mutual Information based labelling and comparing clusters. In Gläser, J., Scharnhorst, A. & Glänzel, W. (eds), Same data – different results? Towards a comparative approach to the identification of thematic structures in science, Special Issue of Scientometrics X(Y):XYZ, doi:10.1007/s00000-000-0000-0

Koopman R, Wang S, Scharnhorst A (2017) Contextualization of topics - Browsing through the universe of bibliographic information. In Gläser, J., Scharnhorst, A. & Glänzel, W. (eds), Same data – different results? Towards a comparative approach to the identification of thematic structures in science, Special Issue of Scientometrics X(Y):XYZ, doi:10.1007/s00000-000-0000-0

Lambiotte R, Rosvall M (2012) Ranking and clustering of nodes in networks with smart teleportation. Physical Review E 85(5):056,107

Mirshahvalad A, Lindholm J, Derlen M, Rosvall M (2012) Significant communities in large sparse networks. PloS one 7(3):e33,721

Möller U (2005) Estimating the number of clusters from distributional results of partitioning a given data set. In: Adaptive and natural computing algorithms, Springer, pp 151–154

Morris S, Van der Veer Martens B (2008) Mapping research specialties. Annual review of information science and technology 42(1):213–295

Newman M (2003) The structure and function of complex networks. SIAM Review 45:167–256

Rosvall M, Bergstrom C (2009) Fast stochastic and recursive search algorithm, preprint available at http://www.tp.umu.se/ rosvall/algorithm.pdf

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

Rosvall M, Bergstrom CT (2010) Mapping change in large networks. PloS one 5(1):e8694

Rosvall M, Bergstrom CT (2011) Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. PloS one 6(4):e18,209

Rosvall M, Axelsson D, Bergstrom CT (2010) The map equation. The European Physical Journal Special Topics 178(1):13–23

Rousseau R (1999) Temporal differences in self-citation rates of scientific journals. Scientometrics 44(3):521–531

Shibata N, Kajikawa Y, Takeda Y, Matsushima K (2009) Comparative study on methods of detecting research fronts using different types of citation. Journal of the American Society for Information Science and Technology 60(3):571–580

Van Eck NJ, Waltman L (2010) Software survey: Vosviewer, a computer program for bibliometric mapping. Scientometrics 84(2):523–538

Van Eck NJ, Waltman L (2017) Citation-based clustering of publications using CitNetExplorer and VOSviewer. In Gläser, J., Scharnhorst, A. & Glänzel, W. (eds), Same data – different results? Towards a comparative approach to the identification of thematic structures in science, Special Issue of Scientometrics X(Y):XYZ, doi:10.1007/s00000-000-0000-0

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, Haque A, Lagoze C (2010) A new approach to analyzing patterns of collaboration in co-authorship networks: mesoscopic analysis and interpretation. Scientometrics 85(1):219–242

Velden T, Haque A, Lagoze C (2011) Resolving author name homonymy to improve resolution of structures in co-author networks. In: JCDL'11, June 13-17, 2011, Ottawa, Ontario, Canada

Velden T, Cambo S, Ahmed S, Lagoze C (2013) Toward a time-sensitive mesoscopic analysis of co-author networks: A case study of two research specialties. In: ISSI 2013, 15-19 July, Vienna, Austria

Velden T, Yan S, Yu K, Lagoze C (2015) Mapping the evolution of scientific community structures in time. In: Proceedings of the 24th International Conference on World Wide Web, ACM, pp 1039–1044

Velden T, Boyack K, Gläser J, Koopman R, Scharnhorst A, Wang S (2017) Comparison of Topic Extraction Approaches and Their Results. In Gläser, J., Scharnhorst, A. & Glänzel, W. (eds), Same data – different results? Towards a comparative approach to the identification of thematic structures in science, Special Issue of Scientometrics X(Y):XYZ, doi:10.1007/s00000-000-0000-0

West JD, Wesley-Smith I, Bergstrom CT (2016) A recommendation system based on hierarchical clustering of an article-level citation network. IEEE Transactions on Big Data (1):1–1

Zuccala A (2006) Author cocitation analysis is to intellectual structure as web colink analysis is to... ? Journal of the American Society for Information Science and Technology 57(11):1486–1501