Scientometrics
https://doi.org/10.1007/s11192-021-03920-3

Challenges to the validity of topic reconstruction

Matthias Held[1] · Grit Laudel[2] · Jochen Gläser[1]

Received: 8 December 2020 / Accepted: 16 February 2021

Abstract
In this paper we utilize an opportunity to construct ground truths for topics in the field of atomic, molecular and optical physics. Our research questions in this paper focus on (i) how to construct a ground truth for topics and (ii) the suitability of common algorithms applied to bibliometric networks to reconstruct these topics. We use the ground truths to test two data models (direct citation and bibliographic coupling) with two algorithms (the Leiden algorithm and the Infomap algorithm). Our results are discomforting: none of the four combinations leads to a consistent reconstruction of the ground truths. No combination of data model and algorithm simultaneously reconstructs all micro-level topics at any resolution level. Meso-level topics are not reconstructed at all. This suggests (a) that we are currently unable to predict which combination of data model, algorithm and parameter setting will adequately reconstruct which (types of) topics, and (b) that a combination of several data models, algorithms and parameter settings appears to be necessary to reconstruct all or most topics in a set of papers.

Keywords Mapping · Topic reconstruction · Research trails · Validity · Infomap · Network clustering

Introduction

Reconstructing research topics from networks of papers is considered a major challenge that keeps attracting attention, and for which new solutions are suggested (Gläser et al., 2017; Held & Velden, 2019; Klavans & Boyack, 2017b; Šubelj et al., 2016). The attention that is paid to this endeavour stems mainly from the interest to identify research areas or 'hot topics' for management or policy purposes. In this context, many researchers have

✉ Matthias Held
matthias.held@tu-berlin.de

Grit Laudel
grit.laudel@tu-berlin.de

Jochen Gläser
jochen.glaeser@tu-berlin.de

[1] Social Studies of Science and Technology, TU Berlin, Hardenbergstr 16-18, 10623 Berlin, Germany

[2] Institute of Sociology, TU Berlin, Fraunhoferstraße 33-36, 10587 Berlin, Germany

Published online: 06 March 2021

acknowledged the necessity to validate the approaches used. If we neglect the validation of approaches to topic reconstruction, we are unable to create knowledge that can be built upon, i.e. we hamper cumulative research and progress in the field:

> So far, approaches to and results of topic identification exercises appear to be incommensurable. This is highly unsatisfying because there is no cumulative growth of knowledge, which means that the local progress made by many bibliometricians does not translate into progress of topic identification as a research area. Progress of a research area is possible only when findings can be related to each other and be placed in a larger consistent framework, which evolves with each contribution that is placed in it. (Gläser et al., 2017: 987).

The validation of topic reconstruction depends on a common standard that represents a ground truth.[1] Klavans and Boyack (2017b: 991) bemoan that "there are no agreed-upon gold standards (defined as examples of ground truth) for literature partitions", a sentiment that was stated earlier by Waltman and van Eck (2012: 2390). In lieu of tests against a ground truth, several strategies have been developed in bibliometrics to tackle the problem of validation. The first strategy utilizes bibliometric surrogates of topics, i.e. bibliometric data that 'stand in' for a ground truth based on a plausible argument (e.g. synthesis papers as 'gold standard', Klavans & Boyack, 2017b; Sjögårde & Ahlgren, 2018).[2] The second strategy makes direct use of expert validation (as e.g. in Chumachenko et al., 2020; Haunschild et al., 2018), and the third strategy compares the results to other, 'trustful' classification systems such as PubMed's Medical Subject Headings, which are created by experts (e.g. Ahlgren et al., 2020).

None of these strategies is satisfactory. Since the relationship between the surrogates and topics remains unknown, none of the strategies can establish the validity of an approach to topic reconstruction. For knowledge in the area of topic reconstruction to grow cumulatively, we need a shared definition of a scientific topic as a basis for the construction of a ground truth *ex ante,* against which the outcomes of topic-reconstruction exercises could be tested. Only then we can comparatively assess the suitability of any approach for the reconstruction of one of the ground truths.

In this paper we utilize an opportunity to construct ground truths for topics in the field of atomic, molecular and optical (AMO) physics. Our research questions in this paper focus on (i) how to construct a ground truth for topics and (ii) the suitability of common algorithms applied on bibliometric networks to reconstruct these topics.

We use the ground truths to test two data models (direct citation and bibliographic coupling) with two algorithms (the Leiden algorithm and the Infomap algorithm), which are popular in the bibliometrics community (Ahlgren et al., 2020; Bohlin et al., 2014; Šubelj et al., 2016; Velden, Boyack, et al., 2017; Velden, Yan, et al., 2017). Our results are discomforting: none of the four combinations leads to a consistent reconstruction of the ground truths. No combination of data model and algorithm simultaneously reconstructs

[1] By 'ground truth' we mean properties of empirical objects that make them instances of the class we created with the definition of a concept, and which we intend to reconstruct. A ground truth for topics thus consists of empirical properties of a set of knowledge claims that makes them a topic according to a definition of the concept. Such ground truths have not yet been used by bibliometrics.

[2] In the context of this strategy, 'validity' is sometimes replaced by 'accuracy' (Ahlgren et al., 2020; Klavans & Boyack, 2011, 2017b). However, this move is merely rhetorical, which becomes obvious when one asks what the authors want to reconstruct with greater accuracy.

all micro-level topics at any resolution level. Meso-level topics are not reconstructed at all. This suggests (a) that we are currently unable to predict which combination of data model, algorithm and parameter setting will adequately reconstruct which topics, and (b) that a combination of several data models, algorithms and parameter settings appears to be necessary to reconstruct all or most topics in a set of papers.

The following sections provide theoretical considerations from which a definition of scientific topics can be derived (2), followed by a general description of our approach as well as the data and methods used (3). We present our results (4) and discuss possible causes for these negative results (5). Conclusions address possible directions of further work (6).

Theory and its consequences

The literature on topic reconstruction neglects its conceptual foundation and leaves its core concept 'topic' undefined. The use of the word 'topic' in publications suggests that a topic is commonly understood as a collection of thematically similar publications (e.g. Klavans & Boyack, 2017b; Šubelj et al., 2016). This implicit definition sees a publication as a container of knowledge that belongs to a particular topic, and the task of topic reconstruction as identifying the correct set of containers. It underlies the task of classification: "A classification system of science assigns journals or individual publications to research areas" (Waltman & van Eck, 2012: 2378, see also Klavans & Boyack, 2017b). Scientific knowledge is assumed to be unambiguously structured, i.e. to consist of relatively stable bounded areas – topics—that are represented by clusters of publications. The validity of a method of topic reconstruction can be understood as the degree to which it reconstructs the clusters of publications that represent topics.

Although bibliometric studies of topic reconstruction do not engage with theories of knowledge production, their implicit understanding of topics as bounded areas of structured knowledge links bibliometrics to science studies, which consider the role of knowledge as input, condition and outcome of practices of knowledge production. A major difference between bibliometrics and science studies is the latter's focus on practices of using and creating knowledge, which emphasizes that all knowledge structures are made by researchers. Researchers create knowledge claims and, in that process, link them to other knowledge claims. There is no scientific knowledge other than that made by researchers producing and connecting knowledge claims. Topics as knowledge structures are a product of a collective enterprise of knowledge production. They order this knowledge production by providing a shared frame of reference from which researchers derive problems to work on and the means of solving these problems.

A sociological definition of the concept 'topic' embeds it in conceptual considerations of the coordinating role of knowledge structures in the production of knowledge. These considerations can be traced back to Fleck (1979 [1935]) and became influential in science studies with Kuhn's (1962) concept of a 'paradigm'. The important idea Kuhn introduced in science studies was that of knowledge (rather than social norms) orienting problem choices and the selection of approaches in science. The work on scientific specialties and fields in the 1970s was based on the idea of an interaction of intellectual and social structures (Chubin, 1976; Edge & Mulkay, 1976; Whitley, 1974, 2000 [1984]). Based on these ideas, a topic can be defined as "*a focus on theoretical, methodological or empirical knowledge that is shared by a number of researchers and thereby provides these researchers with a joint frame of reference for the formulation of problems, the selection of methods* *or objects, the organisation of empirical data, or the interpretation of data"* (Havemann et al., 2017: 1091).[3]

According to this definition, a topic is a *collective interpretation* of a set of knowledge claims that delineates these claims by distinguishing them from knowledge claims that do not belong to this set, connects the knowledge claims in specific ways, and provides avenues for the topic's development by pointing to knowledge that would belong to the topic but has not yet been produced.

On first glance, the definition of topics as collective perspectives on knowledge appears to be an unnecessary and unmanageable theoretical complication. It seems unnecessary because bibliometrics is, after all, interested in knowledge structures rather than perspectives that create them. It seems unmanageable because the collective perspectives cannot be reconstructed with bibliometric methods.

Nevertheless, we see five major advantages of this definition. First, it establishes a link between bibliometrics and those areas of science studies that are interested in topics and their dynamics as a condition of and input in the production of knowledge. This opens a door for collaboration. What appeared to be a difference in interests – the interest in the social construction of knowledge versus the interest in the dynamics of knowledge structures – turns out to be a difference in perspectives on the same empirical object. Both the theoretical grounding of bibliometrics and the use of bibliometric methods for science studies can benefit from this link.

Second, the recalcitrant properties of topics that have been plaguing bibliometrics for decades can be better understood and thus addressed. These include the diffuseness and shifting nature of topics which don't seem to have 'natural' boundaries (Zitt et al., 2005) as well as their varying sizes, their overlap, and their local character. For example, the local character of topics challenges a recently emerging preference for global approaches to topic reconstruction. Klavans and Boyack (2017a: 1160) argued that

> a local model typically excludes many papers that are highly linked to either journal or keyword-based datasets (lower recall) and includes papers that are not well linked within the set and thus belong elsewhere (lower precision). Global models, by definition, include much more of the relevant content.

The implicit premise of this argument is that all content plays the same role in the reconstruction of topics. If this were true, more content would indeed always be better. However, from our definition follows that the relevance of content for a particular topic varies. Content outside the shared focus constituting a topic differs from content that 'belongs' to that topic in its functions and relevance. In other words, it is important to distinguish between an insider perspective on a topic, i.e. the delineation of knowledge by researchers working with it, and an outsider perspective, i.e. the delineation of knowledge by researchers observing it while working on different topics. Local models reconstruct topics from an insider perspective, while global models construct a compromise of both perspectives in which outsider perspectives have a greater influence than the insider perspective. Which approach is preferable depends on the research question.

[3] The definition of a topic as a collectively shared perspective on knowledge positions it in social theory as a specific instance of collective frames, i.e. interpretive schemes that are shared by a group of actors and maintained through their interactions. Interpretive schemes are specific cognitive structures that provide knowledge about situations and proven solutions for typical problems (Schütz, 1967; Schütz & Luckmann, 1973; Giddens, 1979). The term 'frame' was coined by Goffman (1974).

Third, the multiplicity of perspectives also explains the existence of multiple ground truths. Knowledge structures may be delineated differently by different researchers or societal actors depending on their scientific perspectives or interests, which means that "[t]here are multiple partial 'ground truths', which cannot easily serve as a yardstick for topic identification" (Gläser et al., 2017: 987). This statement has been misread as the position that "different approaches give different results that are each meaningful and that the accuracy of different approaches cannot really be established" (Klavans & Boyack, 2017a: 1161) or that "each relatedness measure yields clustering solutions that are accurate in their own right" (Waltman et al., 2020: 691). In contrast, the assumed existence of multiple ground truths only implies that more than one clustering solution *may* be valid because different clustering solutions *may* reconstruct different ground truths. This is a far cry from 'anything goes'. The problem of validating clustering solutions just becomes more complicated because instead of using one 'gold standard', validation requires specifying and justifying *ex ante* which of the several ground truths is to be used, and assessing the clustering solution against it.

Fourth, it becomes clear why expert validation never completely fails or succeeds (Gläser, 2020; see Haunschild et al., 2018 for a recent illustration). Any expert validation provides a perspective on the knowledge structure that partially agrees because the expert belongs to the field but has an individual perspective that does not fully coincide with the perspectives of other experts or with the collective perspective of researchers working with the topic. Furthermore, experts are unlikely to be competent for the whole clustering solution presented to them.

Finally, the definition provides guidance for the construction of ground truths. From the definition of a topic as a shared perspective on knowledge follows that the ground truth has to be constructed from the knowledge of those applying this perspective. We must investigate the perspectives of researchers working on a particular topic in order to deduce their perspectives, to reconstruct their shared focus, and to identify the publications that represent this focus. This approach might be surpassed by others once we have acquired systematic knowledge about the relationship between (types of) topics and bibliographic metadata. Until then, constructing a ground truth for topic reconstruction exercises is a challenge to qualitative science studies, which need to reconstruct the ground truth from observations of and interviews with researchers. The problem for qualitative science studies is that they have only access to researchers' individual perspectives, from which they need to reconstruct collective-level knowledge structures. With this paper, we demonstrate how this might be achieved by integrating bibliometric techniques into qualitative research. We use individual-level and collective-level ground truths for the assessment of clustering solutions.

Approach

Strategy

Topics as collectively shared perspectives on knowledge cannot be used as ground truths for assessing the outcomes of bibliometric topic reconstruction exercises because the former is an interpretive scheme embedded in human consciousness and the latter is a structured set of publications. To overcome this incompatibility, the ground truths must be constructed as sets of publications that validly represent topics. These

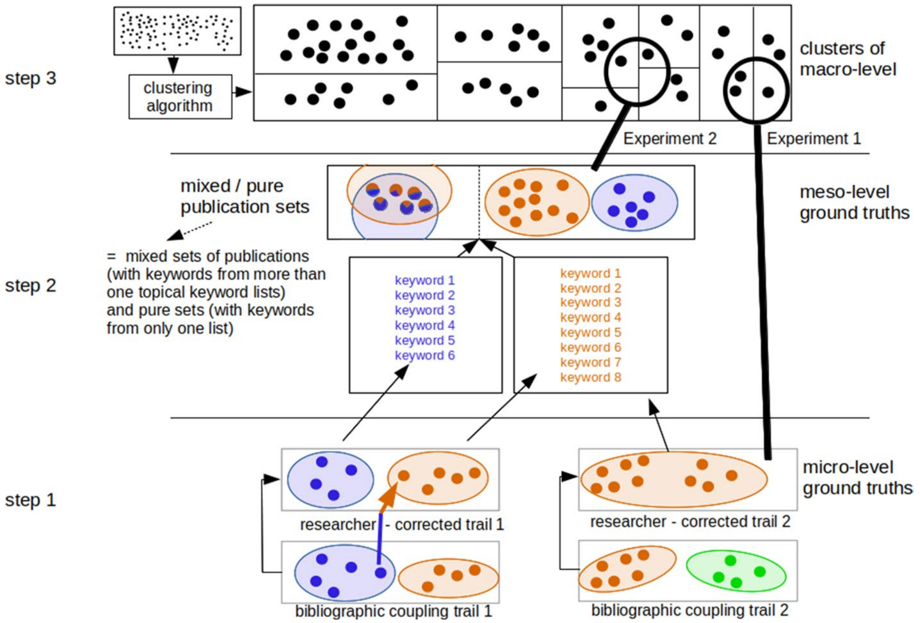

step 3
clustering algorithm
clusters of macro-level
Experiment 2
Experiment 1
mixed / pure publication sets
= mixed sets of publications (with keywords from more than one topical keyword lists) and pure sets (with keywords from only one list)
meso-level ground truths
step 2
keyword 1
keyword 2
keyword 3
keyword 4
keyword 5
keyword 6
keyword 1
keyword 2
keyword 3
keyword 4
keyword 5
keyword 6
keyword 7
keyword 8
step 1
researcher - corrected trail 1
researcher - corrected trail 2
micro-level ground truths
bibliographic coupling trail 1
bibliographic coupling trail 2

Fig. 1 Strategy for using ground truths to evaluate clustering solutions

ground truths need to be derived from researchers' perspectives. Therefore, we utilize a bottom-up strategy for the construction of ground truths (Fig. 1). We first obtain individual researchers' perspectives on sets of their own publications, which represent their 'research trails', i.e. the sequence of topics they have worked on (step 1). These 'elementary topics' of researchers have all important properties of a topic as defined in the previous section except one. What researchers consider as 'their topics' are *foci on theoretical, methodological or empirical knowledge* that serve as frames of reference and guide individual research. They are *not*, however, *collectively shared* perspectives. Thus, we can consider them as elementary models of topics but not as topics according to our definition. They can serve to construct our micro-level ground truths because these ground truths should be easier to reconstruct than those derived from 'true' topics, which only exist on the meso-level.

In step 2 we construct meso-level ground truths by integrating the publications researchers assigned to their individual topics with a keyword-based approach. This process leads to two types of meso-level ground truths, namely 'pure' publication sets that represent only one topic and 'mixed' publication sets that represent several overlapping topics. In step 3 we create a set of publications representing AMO research, create two data models (direct citation and bibliographic coupling), and cluster the networks.

With these clustering solutions we conduct two experiments. In experiment 1, we test whether the four combinations of data models and algorithms can reconstruct our individual-level ground truths, i.e. the sets of publications that were described as belonging to the same topic by researchers in interviews. In experiment 2, we test whether the four combinations can reconstruct the meso-level ground truths. We select the easiest task for the algorithms and only evaluate the reconstruction of the 'pure' meso-level ground publication sets, i.e. those publication sets that have only keywords from one topic.

Step 1 and 2: Constructing Ground Truths

Data

For the construction of the micro-level ground truths (step 1), we used data from a project on the emergence of experimental BEC in the 1990ies and 2000s (Laudel et al., 2014). In this project, 38 researchers working in atomic and molecular optics (AMO) who switched or did not switch topics by turning to experimental Bose–Einstein condensation (BEC) at different points in their careers were interviewed. Bose–Einstein condensates occur when gases of atoms are cooled to temperatures very close to absolute zero, particles lose their individual identities and coalesce into a single blob. Their existence was theoretically predicted in 1924 by Bose and Einstein. Although BEC was widely accepted as a theoretical possibility, its experimental realisation was regarded by many physicists as very difficult, if not impossible, to achieve for both theoretical and technological reasons. In 1995 the first BEC was experimentally produced by two US groups, an achievement that was later rewarded with the Nobel Prize. The scientific community was initially undecided whether BEC would be the end of a long quest or whether it would open up new research opportunities. However, it soon recognised that BEC can be used for a wide range of fundamental research in several subfields of physics and BEC research grew rapidly to become an established field of research (Fallani & Kastberg, 2015).

For the construction of micro-level and meso-level ground truths, we used interviews with twelve researchers who switched and two researchers who did not switch topics by turning to experimental BEC. By including AMO researchers who switched to experimental BEC, we can further specify one of our ground truths – publications on experimental BEC began in 1995 – and can ask whether clustering solutions reconstruct the emergence of experimental BEC.

Prior to the interviews with these AMO researchers, their publication metadata were downloaded from the Web of Science (WoS), and publication clusters representing their research biographies were constructed. These bibliometric data were discussed in interviews. Researchers confirmed or corrected the correspondence of clusters of their publications with topics they worked on. In the present study, we used data from fourteen of these AMO physicists.

The construction of the meso-level ground truths (step 2) was based on the results from step 1, namely the clusters of researchers' publications that represented topics researchers worked on. From these publications, we extracted specific keywords and created topical keyword lists.

Methods

In the project conducted by Laudel et al. (2014), face-to-face semi-structured interviews were used. A major thematic focus of the interviews were the interviewee's research topics beginning with their PhD topic, with an emphasis on thematic changes and the reasons for them. In addition, developments in the interviewee's national and international communities were discussed. This part of the interview centred on graphical representations of the interviewees' research trails. Research trails were constructed by downloading their publications from the Web of Science, constructing bibliographic coupling networks (using Salton's cosine for bibliographic coupling strength) and choosing a threshold for the strength

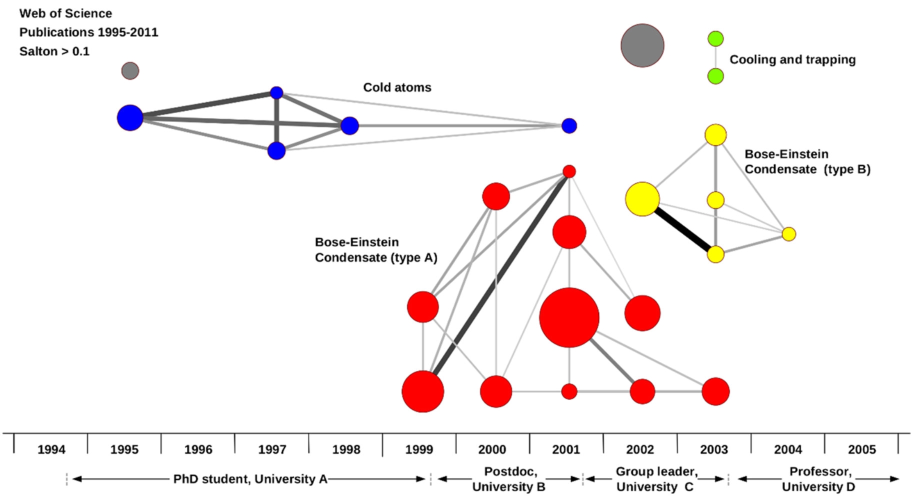


Fig. 2 Graphic representation of the cognitive career of one of the AMO physicists

of bibliographic coupling at which the network disaggregates into components (Gläser & Laudel, 2015). Although this 'manual' approach also produces several unassigned publications, it is preferable to algorithmic clustering for the construction of an input to interviews. The visual representations of research trails serve as means of 'graphic solicitation' in interviews, for which instant visual recognition of different topics is essential. The components are intended to provide a 'draft' of topics a researcher has worked on over time (see Fig. 2 for an example). Their visualisations were used to prompt narratives about the content of the research at the beginning of the interview (for an extended description of the approach see Gläser & Laudel, 2015). During these narratives, researchers confirmed and sometimes corrected the picture by combining or separating clusters because they perceived research topics as belonging together or being separate (see the micro-level ground truths layer in Fig. 1). The interviews lasted on average 90 min and were fully transcribed. Transcripts were analysed by qualitative content analysis, i.e. we extracted relevant information from the transcripts by assigning it to categories that were derived from our conceptual framework (Gläser & Laudel, 2013, 2019).

There was no straightforward way in which individual-level ground truths – the sets of publications researchers identified as representing one of the topics they worked on—could be aggregated to meso-level ground truths.

As was to be expected, the perspectives of researchers on their topics differed. For example, central methods used by AMO physicists were considered a separate topic by some researchers, while others integrated them in the theoretical topics they worked on. Furthermore, interviewees assigned publications to disjunct topics while 'true' meso-level topics overlap. This is why simply combining all publications from the 14 researchers and applying the method for the construction of research trails to the combined set of publications did not produce any meaningful results, i.e. no threshold could be found at which components represented researchers' topics. We tested this with the publications on the BEC topic and found that setting a low threshold led to BEC being combined with other topics in one component, while increasing the threshold made the component disintegrate into sub-graphs that did not represent researchers' topics.

Table 1 List of the four ‘pure’ topics with the recall that is achieved by searching for the specific keywords (right column) in publications’ metadata

‘Pure’ topic (recall)	Keywords
Bose–einstein condensation (92%)	Magnetic trap, relaxation rates, sodium, bose einstein condensation, bose, spin polarized hydrogen, condensates, evaporative cooling, photoassociation, magneto optical trap
Cooling (80%)	Cooling, cold, interaction potentials, photon recoil, m/s
Precision spectroscopy (91%)	Rydberg, spectroscopy, pulse
Atom interferometry (88%)	Interferometry, diffraction

We therefore used the information from the interviews and a keyword-based approach to construct meso-level ground truths:

1. Information from interviews was used to distinguish between ‘pure’ individual topics – those that addressed only one set of problems—and ‘mixed’ topics – those combining topics that were considered separate by other researchers.
2. Publications assigned by the researchers to ‘pure’ topics and contained in the AMO dataset described below ("Step 3: Obtaining Clustering Solutions"section.) were aggregated. This resulted in four sets of more than ten publications, which were used for the extraction of keywords. These included Bose–Einstein condensation (71 publications), cooling (25), atom interferometry (24) and precision spectroscopy (22).
3. Keywords were extracted from all publications linked to each topic. We used author keywords, keywords plus, and the terms which we extracted from titles and abstracts. The term extraction was conducted with a rule-based approach using a noun phrase chunker in python’s ‘nltk’ package.[4] Based on the information from the interviews, groups of semantically equivalent terms could be identified. For example, *laser cooled rubidium* and *ultracold rubidium* could be subsumed to *rubidium* because all researchers worked at ultra-low temperatures, and laser cooling was the only method to achieve these temperatures.
4. In order to determine sets of topic-specific terms, we treated each publication linked to one of the four ‘pure’ topics as a cluster and calculated the normalized mutual information (NMI) value for each keyword (Koopman & Wang, 2017). Due to the large differences in the sizes of the topics the results also included very general, frequently occurring terms. We manually excluded these terms (e.g. ‘atoms’, ‘gas’) from the lists of specific terms for each topic.
5. We ranked terms according to their NMI values and iteratively selected terms for each topic until the set of terms allowed us to retrieve most of the publications of each topic. This led to a minimum set of keywords which occur in a large proportion of the publications linked to a topic, and occur mostly in these publications, hence can be considered specific for this topic (Table 1).

[4] Python code of the script used is available at https://github.com/insi01/noun_phrase_chunker.

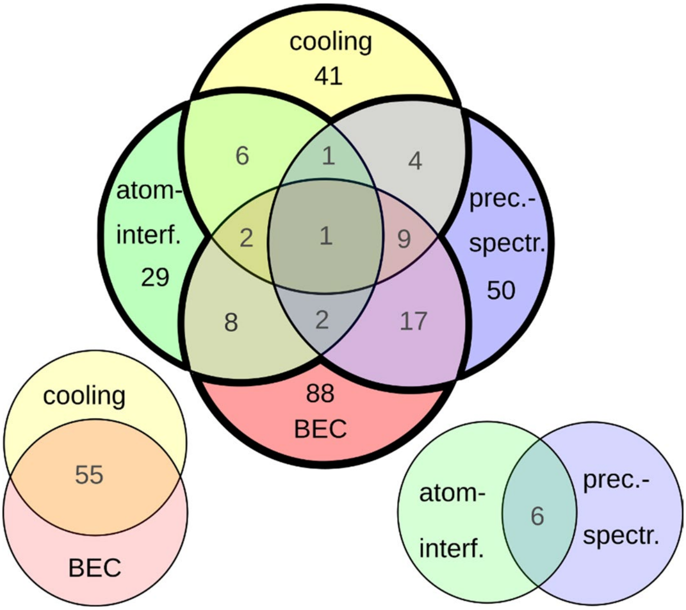


Fig. 3 Venn diagrams of overlaps between the four sets of publications with numbers of publications

6. These keyword lists were applied to retrieve researchers' publications. Not surprisingly, many publications contained keywords from more than one list and were thus linked to more than one topic.

This way, the disjunct assignment of publications by interviewees could be replaced with sets of publications in which topics overlap (Fig. 3). The areas bounded by bold lines in Fig. 3 represent 'pure' parts of ground truths, i.e. sets of publications that contain keywords from only one topical list. Reconstructing them can be considered the easiest task, which is why we used only these 'pure' meso-level ground truths in our experiments.

Step 3: Obtaining Clustering Solutions

Data

To construct the macro-level AMO dataset, we started delimiting the Web of Science by selecting all publications from journals in the subject category 'Physics, Atomic, Molecular & Chemical' published 1975–2005, excluding physical chemistry journals by searching for 'chemi' in the journal titles. We then expanded and refined this dataset by:

1. Including publications from all other physics subject categories from 1990–2005 that cited at least two publications from the initial data set;
2. Limiting the data set to publications from 1990 to 2005; and
3. By including publications from all other physics subject categories from 1990 – 2005 that have been co-cited with at least two papers from the 1990—2005 data set.

This resulted in 369,188 publications, whose metadata were downloaded from the 2018 stable version of the Web of Science database hosted by the 'Kompetenzzentrum Bibliometrie' (KB[5]). From this dataset we created a subset of publications that contained all AMO physics research relevant for our 14 researchers. To achieve this reduction, we decided to build and cluster the direct citation network[6] of this dataset, with the giant component containing 366,480 publications, which included all relevant publications of the micro-level research trails. We applied the Leiden algorithm (Traag et al., 2018) for a coarse clustering (resolution 8e-6) of the giant component, and then extracted the largest cluster with 96,137 publications including 415 (78%) of the research trails' publications. The missing 22% of the researchers' publications contain mainly research not belonging to AMO physics or research at the borders of the field (e.g. interdisciplinary collaborations, application of common AMO physics methods in other fields). These 96,137 publications served as our macro-level AMO dataset. For this dataset, two data models were created.

The direct citation network was created using only the 'internal' links, i.e. only citations from and to publications in the AMO dataset. Weights were attached to the links according to the normalization formula in (Waltman & van Eck, 2012: 2380).[7] For constructing the bibliographic coupling network, only references have been used which are source items in the Web of Science. Here, weights were attached to the links by calculating the Salton's cosine value.

Methods

In order to detect communities in both networks, we selected two algorithms popular in the scientometrics community, namely the so-called Leiden algorithm (v. 1.0.0) (Traag et al., 2018), which further develops the widely used Louvain algorithm, and the Infomap algorithm (v. 0.21.0) (Rosvall & Bergstrom, 2008), which has also already successfully been applied in bibliometrics (Šubelj et al., 2016; Velden, Boyack, et al., 2017; Velden, Yan, et al., 2017).

Both algorithms allow for parameter settings which create coarser or more fine-grained solutions. We varied this parameter in order to allow both algorithms to reconstruct smaller or larger topics in our dataset. The Leiden algorithm requires the specification of a resolution parameter. This parameter is included in its quality function, the Constant Potts Model (CPM), whose optimization for the chosen resolution value results in a particular partition of the network. The seed parameter was always set to '0′ for all runs. Varying the resolution parameter leads to different numbers of clusters in a partition.

Infomap finds the minimum description length of a random walk in a given network by creating modules. The parameter Markov random time has the standard value of 1. Changing this parameter means changing the number of steps of the random walker which are encoded (Kheirkhahzadeh et al., 2016). This will result in a more or less fine-grained solution. Furthermore, Infomap allows for 'multilevel compression', which means in the result that the 'modular' organization of a network can then be analyzed on several levels

[5] Competence Centre for Bibliometrics. http://www.forschungsinfo.de/Bibliometrie/en/index.php?id=home.

[6] We chose the direct citation model here because it is a popular choice in bibliometrics for the production of macro-level science maps (Shibata et al., 2009; Velden et al., 2017), given their relative sparseness and thus easy computability.

[7] We did not perform any test as to the influence of other weight assignment strategies.

Table 2 Minimum and maximum granularity levels for Leiden algorithm (resolution parameter) and Infomap (Markov random time), as well as steps taken (clustering solutions produced) from minimum to maximum

Granularity level	Leiden DC	Leiden BC	Infomap DC	Infomap BC
Min	6e–06	5e–05	0.8	0.1
Max	7.53e–05	9.95e–03	4.0	1.4
Steps	100	100	33	14

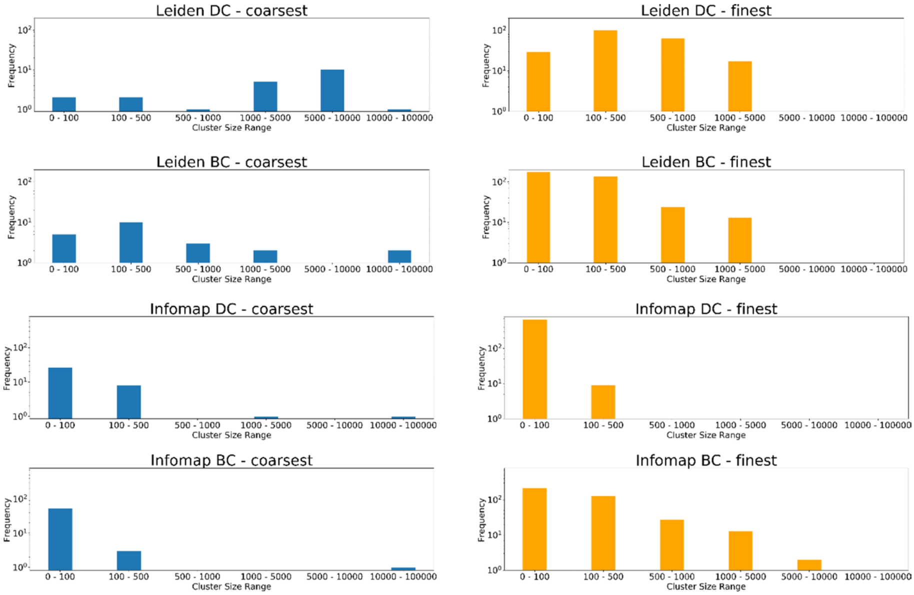


Fig. 4 Cluster sizes distributions for coarsest and finest solution for the cluster solutions

of hierarchy. Here, we decided to always analyze the second lowest level of the hierarchy,[8] since this seemed to be a fair decision considering the sizes of our topics (the lowest level of the hierarchy are individual nodes, the highest level of the hierarchy represents a rather coarse clustering).

The ‘granularity’ intervals were varied by changing the resolution parameter of the Leiden algorithm (higher values lead to more granular solutions) and Markov random time of the Infomap algorithm (higher values lead to less granular solutions). To account for the possibility to find a certain number of topics each of a certain minimum size, each parameter range was chosen to obtain a result with at least 20 clusters that have at least 40 publications each. This constraint sets a lower and an upper bound for the resulting granularity level of the clustering results. Clusters smaller than 20 publications have not been regarded in our analysis because the data set covers 16 years and we assume a topic to be represented by several publications per year. In Table 2 the upper and lower bounds of

[8] Note that the multilevel compression can also be forced to be a two-level compression via a parameter.

the parameter settings are shown, and the resulting size distributions of the coarsest (least granular) and finest (most granular) solution of the four combinations are displayed in Fig. 4. For the applications of the Leiden algorithm to the data models—solutions Leiden DC (direct citation model) and Leiden BC (bibliographic coupling model)—100 solutions each were produced (100 steps from lowest to highest resolution). For Infomap DC and Infomap BC, we produced 33 and 14 solutions, respectively.

Experiments 1 and 2: assessing the clustering solutions

In order to evaluate whether an individual researcher's research topic has been reconstructed by the macro-level clustering, we took the publications that were assigned to a topic by the researchers and searched for them in the clusters produced by the four combinations of data models and algorithms for all parameter settings. To qualify as a successful reconstruction of a ground truth, a cluster had to fulfil two criteria:

1. All publications of a topic belong to the same cluster, and
2. No publications from the researcher's other topics belong to this cluster.

If the publications of a ground truth fall into one cluster, and no publications from other ground truths are in the cluster, the evaluation gets the value 0 (=correct reconstruction). If several ground truths fall into one cluster, it is assigned the value 1 (ground truths lumping). If the publications of one ground truth spread over several clusters, it is assigned the value -1 (ground truth spread).

This assessment was applied to all individual-level ground truths and to the 'pure' meso-level ground truths. In addition, we checked whether clustering solutions could reconstruct the emergence of experimental BEC by analysing the dynamics of clusters in which individual-level and meso-level BEC publication sets were placed.

Results

We detect only minor and no systematic or interesting differences between the results produced by the four combinations of algorithms and data models. In general, the results of the evaluation of experiments 1 and 2 show that the successful reconstruction of a ground truth was a very rare event.

Micro-level reconstruction

The results of the evaluation of the micro-level ground truths reconstruction can be summarized as follows.

1. Only few clustering solutions reconstruct any of the micro-level ground truths. The two other cases occur much more frequently, i.e. ground truths are lumped together in clusters or spread across clusters.
2. Many micro-level ground truths are not reconstructed at any granularity level across the span.

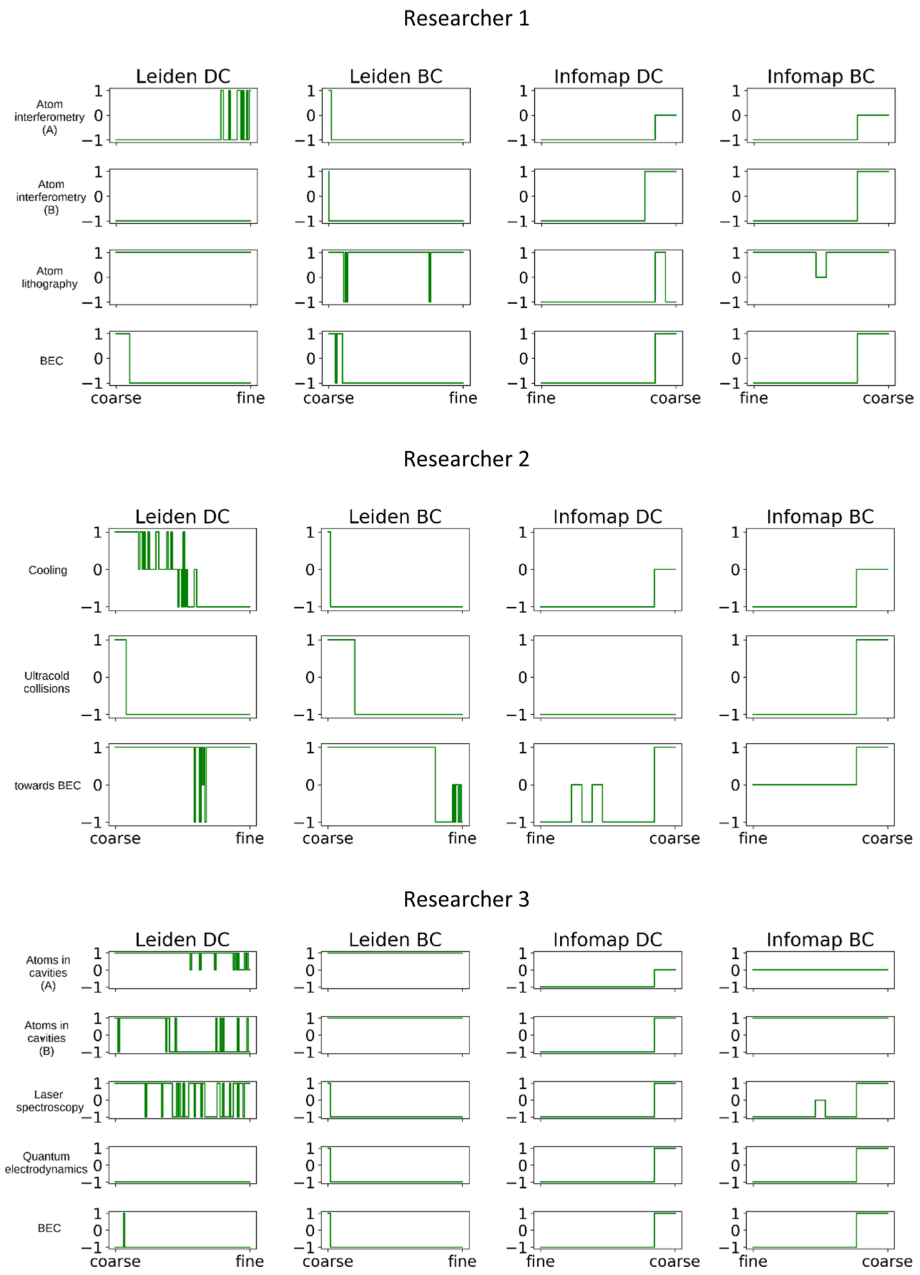


Fig. 5 Reconstruction of ground truths derived from three researchers' individual perspectives on topics by two algorithms applied to two data models (0: ground truth is reconstructed, -1: ground truth is distributed over several clusters, 1: ground truth is combined with other ground truths in one cluster). Note that the x-axis for Infomap is reversed

3. In many cases, micro-level ground truths become split and reunited with changing resolution levels. These fluctuations are not consistent. Counter-intuitively, ground truths can be split at lower granularity and became reunited at higher levels of granularity.
4. No clustering solution reconstructs all micro-level ground truths of a researcher.

Figure 5 displays the evaluation results for three of the twelve researchers who switched to BEC during their career and the publication sets representing their topics. Each image shows at which level of granularity the ground truths were reconstructed, were distributed over different clusters or were combined with other ground truths in one cluster by the specific combination of data model and algorithm. Each row corresponds to one of the researcher's topics and the four columns represent the results of the clustering solutions at different granularity levels. The x-axes correspond to the granularity levels also shown in Table 2 from coarsest to finest.

The results for all three researchers have in common that the assignments of topics to clusters are highly inconsistent. Results for the other nine researchers are not different.

In Fig. 6 the research trail of researcher 1 is plotted onto two selected clustering solutions. We have seen in Fig. 5 that no solution reproduces all topics from researcher one, so it is not surprising to find this also to be the case in selected solutions shown here. However, some patterns can be discerned. Each selected solution has a cluster with publications from all topics of the researcher (lumping). And some publications, e.g. the three triangles of researcher topic 4 around 1996, fall always into one cluster. These patterns can also be discerned from Figure S.3 in the appendix, where other results of the projected research trail of researcher 1 are shown for other clustering solutions.

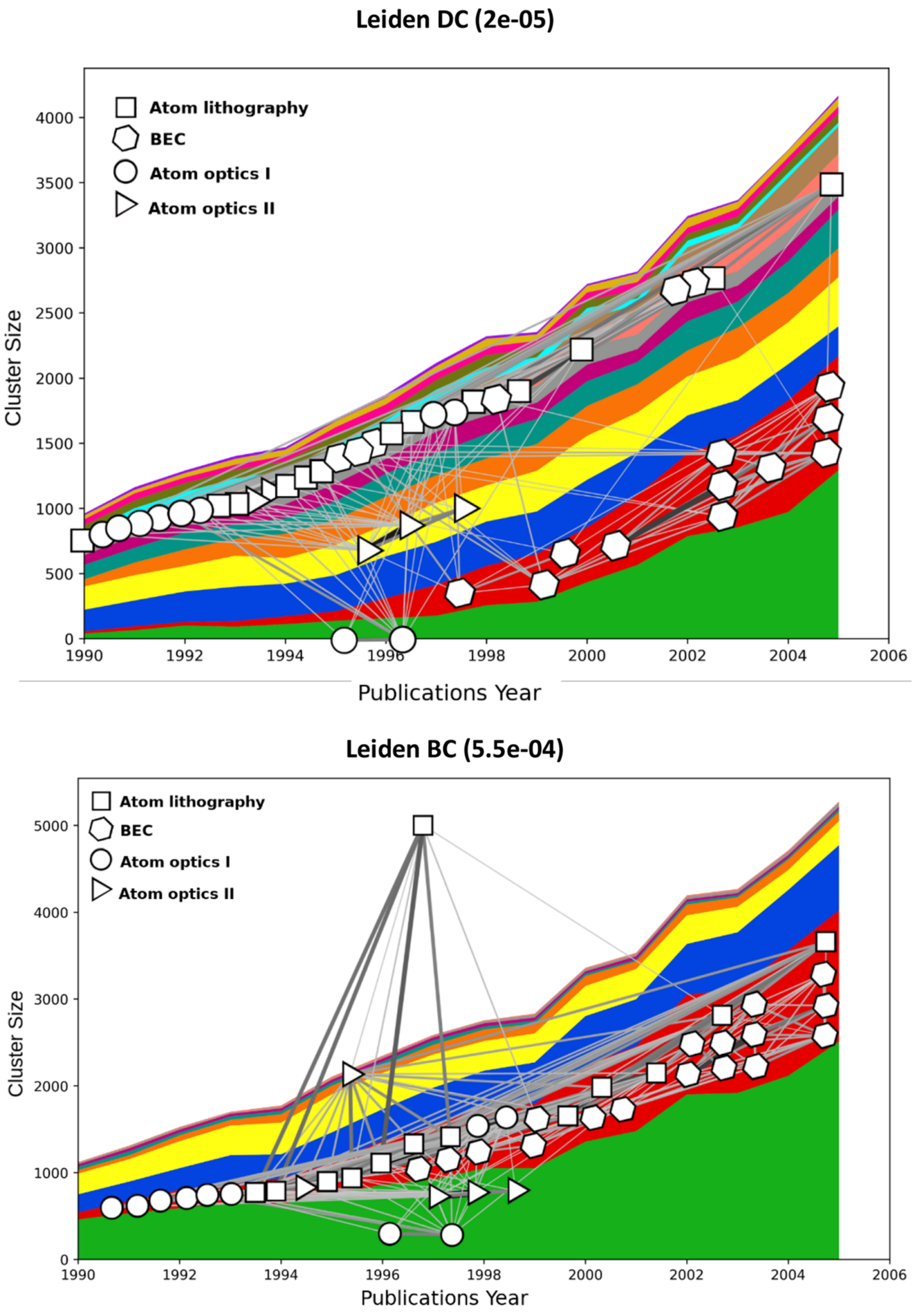
Meso-level reconstruction

The results of the evaluation of the meso-level ground truths reconstruction can be summarized as follows.

1. No solution reconstructs the ground truths for BEC, cooling or precision spectroscopy.
2. Only the smallest ground truth, atom interferometry, is reconstructed in very rare cases.
3. In most cases, meso-level ground truths are spread over several clusters.

Figure 7 displays the evaluation results of experiment 2. The rows represent the different meso-level ground truths. General patterns that we can observe include the wide spread of the ground truth 'precision spectroscopy' over many clusters (across almost all clustering solutions), whereas the topic of BEC is spread over only very few clusters in most cases.

Since in almost all clustering solutions all four meso-level topics are spread over several clusters, we give a more detailed report on selected results for some granularity levels of the two algorithms combined with the two data models. Figure 8 (Infomap DC, an example of a rather coarse solution) shows how the publications of the four meso-level ground truths are placed in the Infomap DC solution with Markov random time 1.5. Note that the square-marked publications each fall into one of the rather small, hardly visible clusters. The unassigned publications on top could not be considered, since they fall into very small clusters with fewer than 20 publications, which were not considered to represent topics. This solution is similar to the coarse solution of Leiden DC in Figure S.1 in the appendix in that it places most of the BEC publications in one cluster, which steadily grows in size over the years. Both solutions, however, also place other publications from the other



Leiden DC (2e-05)

Leiden BC (5.5e-04)

Fig. 6 Projection of the research trail of the four topics of researcher 1 onto two different clustering solutions (the title reads: algorithm, data model, resolution parameter). The width of network edges corresponds to the Salton value of the research trail's bibliographic coupling network

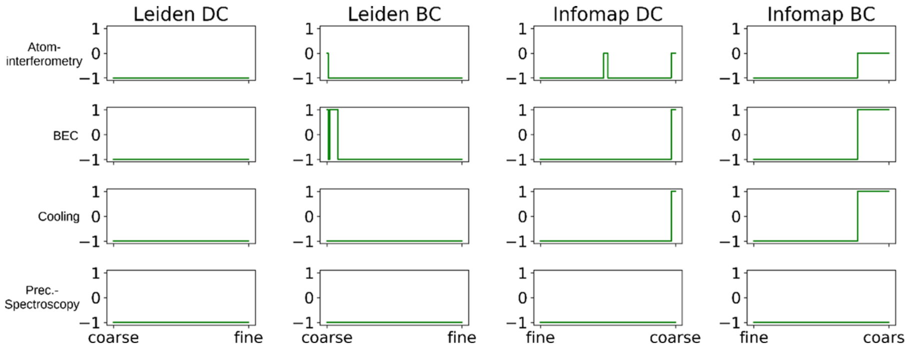


Fig. 7 Reconstruction results (experiment 2) of the meso-level ground truths (0: ground truth is reconstructed, -1: ground truth is distributed over several clusters, 1: ground truth is combined with other topics in one cluster). Note that the x-axis for Infomap is reversed

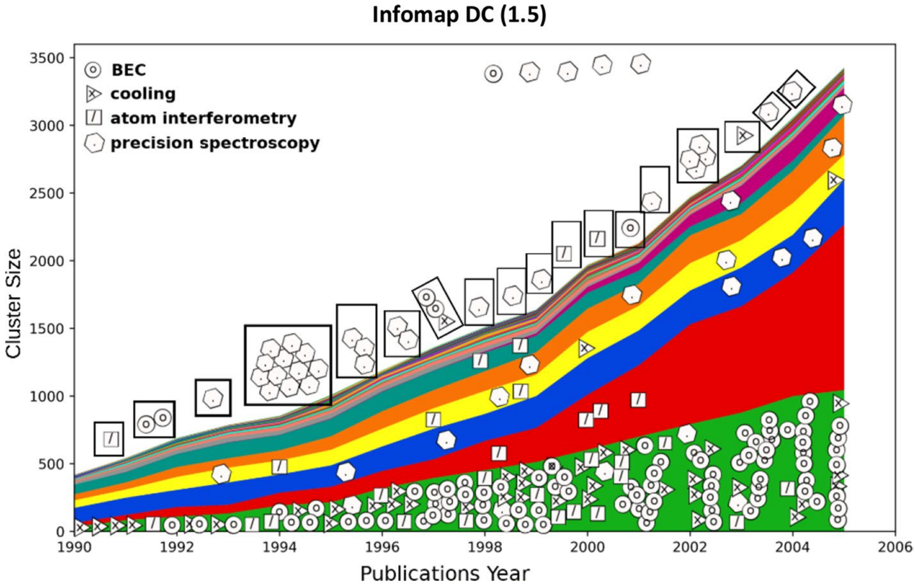


Fig. 8 Projection of the publications of each meso-level topic onto a coarse clustering solution of Infomap DC. Publications in squares and rectangles fall into one of the small, hardly visible clusters

topics in this cluster. Most of the cooling publications and some publications of the other meso-level ground truths are also in this "BEC cluster". For the ground truths representing atom interferometry and precision spectroscopy, no clear corresponding clusters could be identified.

This is also the case in the rather fine-grained solution in Fig. 9, which displays the solution of Leiden DC with resolution 2e-05. There is a large spread of these ground truths over several clusters. Especially precision spectroscopy publications are found in many clusters. In Fig. 9, two clusters together contain almost all of the BEC publications. The bigger, lower of the two clusters contains almost only publications from the BEC ground

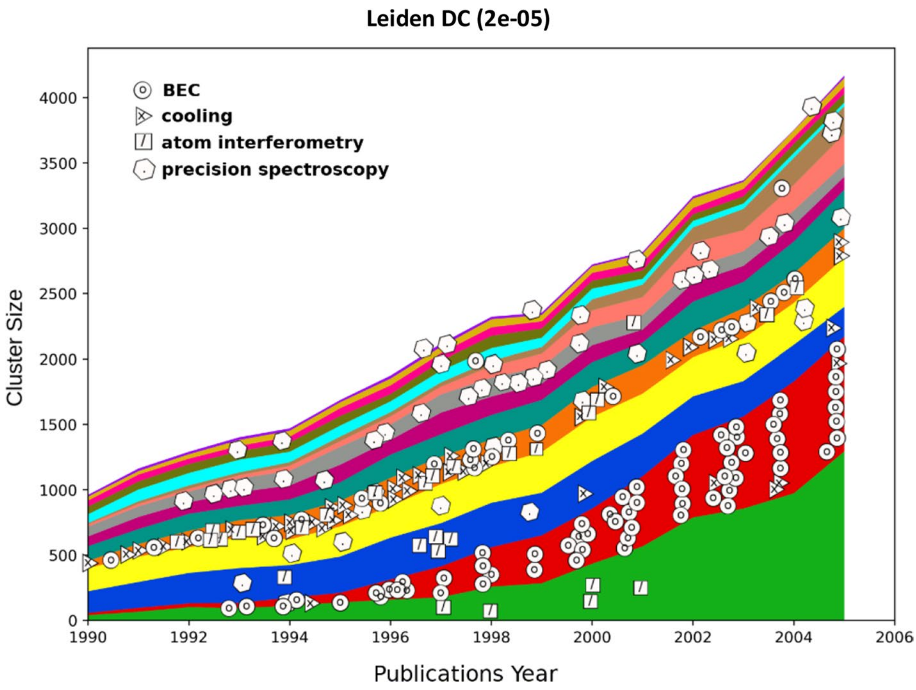


Fig. 9 Projection of the publications of each meso-level topic onto a rather fine-grained clustering solution of Leiden DC

truth together with only few cooling publications. The upper cluster is more mixed but the non-BEC publications are also mostly cooling publications.

Regarding the reconstruction of the emergence of experimental BEC in 1995, we find that there are solutions where clusters markedly 'grow in size' over time, even clusters containing many BEC publications which begin markedly to 'grow' around 1995 (e.g. two solutions in Supplementary Figs. S2). In these 'BEC clusters', however, publications theorizing BEC and those that experimentally realize a BEC are mixed together. Furthermore, almost all clusters in all solutions span the entire time period, i.e. have publications from all years in them. Thus, the emergence of experimental BEC in 1995 could not be reconstructed by any of the clustering solutions.

Discussion

Bibliometric approaches to topic reconstruction are based on one data model only, to which one algorithm is applied whose parameters are changed until it produces a 'satisfactory' solution (as e.g. in Glänzel & Thijs, 2017; Klavans & Boyack, 2017b; Sjögårde & Ahlgren, 2018). The same applies to experiments that systematically compare data models or algorithms (Gläser et al., 2017; Šubelj et al., 2016; Velden, Boyack, et al., 2017; Velden, Yan, et al., 2017). Whatever the data model or algorithm: The outcome of a topic reconstruction exercise is always a clustering solution that is produced by applying one algorithm with one particular setting of parameters to one data model. The implicit assumption underlying this uniform strategy might be that we just have to find the 'right' combination of data model and algorithm to obtain an accurate reconstruction of topics.

In this paper, we applied the same approach but used four combinations of data models and algorithms and a wide range of parameter settings. Still, none of the combinations was able to reconstruct any of the ground truths. Thus, if we apply an algorithm to a data model, we cannot know which ground truth, if any, it reconstructs.

After conducting a preliminary assessment to find reasons on the micro level for the very rare reconstruction of a researcher's topic, we found no correlation to (a) a weak bibliographic coupling between the publications, (b) a unusually high number of citations to publications, or (c) the content of topics (the topics being methods or theories).

In the following, we discuss other possible explanations for the failure of the four approaches to reconstruct our ground truths.

Misconstruction of ground truths

A first possible cause of the disappointing results is that the ground truths we constructed were artefacts rather than ground truths. At the micro-level this could happen for two reasons. First, individual perspectives may be idiosyncratic. When analysing the assignment of publications to topics by interviewees, we observed differences between individual perspectives. For example, one researcher perceives a publication set of his research trail to form a topic, and another researcher perceives thematically very similar publications as belonging to two different topics. This possibility can be directly derived from our definition of topics as shared *perspectives* on knowledge – not all individual perspectives are likely to be shared. Second, our method for obtaining individual perspectives enforced the creation of disjunct clusters of publications, which contradicts the expectation that topics overlap. Since topics emerge from the overlap of individual perceptions, sets of publications that are identified by researchers as belonging to only one of their topics might not be suitable ground truths for topics at all.

While the use of individual perspectives of researchers and of disjunct clusters of publications is at odds with our definition of a topic, it is extremely unlikely that none of the 46 publication sets identified by researchers should match a topic of AMO physics. Regarding the topic of BEC, we are quite sure that the 12 researchers together not only represent a more-or-less collective view on at least one shared topic (BEC), but also share their perspective on topics with at least some other authors in our sample. Even allowing for idiosyncrasies, the small numbers of publications on the micro level belong together in *some* scientific perspectives, which makes it very difficult to believe that the failure of all clustering solutions to reconstruct them across all 12 AMO researchers who started BEC research in the 90s is caused by the way in which these small numbers of publications are categorised by their authors. The individual perspectives on the same topics might not be entirely congruent, but the inconsistency in the results for all 12 perspectives appears to be significant. The individual perspectives of the twelve researchers represent ground truths informed by 'elementary forms' of topics, and it should be possible to reconstruct them even though they do not represent full topics. However, ground truths were not reconstructed in any consistent manner, and it is not predictable which ground truth is reconstructed by any of the four data model/algorithm combinations at one of the granularity levels.

A misconstruction of the meso-level ground truths could occur because we started from individual perspectives despite our premise that topics emerge from a collective perspective, and because we used bibliometric tools to construct the collective perspective. Given the current state of the art, we cannot think of another approach to get an insight into the topics that shape the work of researchers and are shaped by this work. The construction of perspectives and their impact on research starts at the individual level and must therefore be investigated at that level. The collective level of researchers sharing a scientific perspective can be investigated only through an analysis of these researchers' communication through publications, i.e. by using the very means whose validity we want to establish. Nevertheless, our approach still differs from the use of bibliometric 'gold standards' in that it starts from researchers' perceptions of their practices. While bibliometric methods are necessary (and invaluable) for treating epistemic meso-level phenomena, they need firm ground if they are to serve in the construction of ground truths. This firm ground is provided by micro-level ground truths established in interviews with researchers about the content of their research.

For constructing meso-level ground truths, we resolved to overcome the idiosyncrasies of our micro-level ground truths by determining what is common to the individual perspectives – here operationalised as shared, specific keywords – and by identifying these keywords in publications, which created an overlap of topics in publications. This appears to be a viable approach to constructing collective perspectives on knowledge from individual-level data.

This approach, however, is not without difficulties. First, the method we used to determine keywords specific to topics (NMI) does not work well if clusters of publications are of very different sizes. A compromise between keywords that are too specific and keywords that are not specific enough has to be found, and no single solution might be considered best.

Despite these problems, a keyword-based approach to the construction of overlapping topics from individual-level ground truths seems viable because specific keywords often occur in titles, abstracts and keyword lists. Once their specificity and importance for the topic can be determined, they might be used to subsume publications to overlapping topics. The potential of this approach needs to be further explored.

The process of constructing meso-level ground truths is susceptible to inaccuracies. However, the complete failure of all combinations of data models, algorithms and parameters in reconstructing meso-level ground truths cannot be explained by possible (and unavoidable) inaccuracies in the construction of ground truths.

Data models unrelated to topics as shared perspectives

The next two possible explanations for our negative results concern the data models used. These data models could be unsuitable for two reasons. First, the properties of the data model might not represent relevant properties of topics. This question has not yet occurred in bibliometrics because previous studies on the validity of topic reconstruction did not use ground truths. Bibliographic coupling is generally accepted as representing the thematic similarity of two publications (Glänzel & Czerwon, 1996). A direct citation is considered as representing a thematic link between the citing and the cited publication, which is not necessarily an indicator of thematic similarity. Nevertheless, direct citation is still often used to group publications thematically, i.e. used the same way as bibliographic coupling or co-citation, if only the data set is large enough (Boyack & Klavans, 2010; Haunschild et al., 2018; Klavans & Boyack, 2017b; Shibata et al., 2009).

Regarding the weights we attached to the edges of both the direct citation (DC) and bibliographic coupling (BC) network, we are not aware of any studies taking a closer look at the effects of different weight choices on the reconstructed bibliometric structures. Using Salton's cosine for the BC network appears reasonable because it is a plausible way to account for different lengths of reference lists. However, we cannot provide any insight into the effects of this particular choice on the results. We followed the established practice of using the normalization function for calculating the edge weights in the DC model. Again, we are not sure about the normalization's possible adverse effects. Such effects might result from the normalization erasing properties of topics that might be expressed by different citation densities in regions of the network.

Beyond these differences between fundamental properties of the data models, we know nothing about properties of topics and their representation in patterns of referencing. If there are different types of topics which differ in their referencing patterns, each bibliometric data model might be suitable for the identification of some types of topics but not others. These questions are related to the general discussion in network science of reasons why communities found in the metadata structure of a network might not have much to do with the ground truth that one might be looking for (see e.g. Hric et al., 2014; Peel et al., 2017). Utilizing only the topology of a network for community detection might not suffice to reconstruct an underlying ground truth. Various approaches to enrich the topology have been investigated already in network science (for a recent review see Interdonato et al., 2019); and combinations of data models have already been proposed in bibliometrics (e.g. Thijs & Glänzel, 2018). Which features are to be included in data models to improve the reconstruction of topics requires further exploration.

Another aspect that likely has a marked influence on the possibility to reconstruct topics with bibliometric data is the aspect of distortion by long time periods in the data models. Both data models included publications from a time span of 16 years. This means the clustering algorithms applied to the 16-year bibliographic coupling or direct citation network are forced to construct clusters from publications of this whole period. Analysing the impact of different time periods in the data models used on the reconstruction of ground truths was beyond the current study and must be left to future work. One possible avenue could be to explore temporal patterns of the development of a bibliometric network (Small, 2006; Small et al., 2014). This exploration could benefit from insights gained in network science under the heading 'temporal networks' (Interdonato et al., 2019), e.g. by tracing communities over time (Cherifi et al., 2019).

The use of long time intervals is particularly problematic because scientific communities are known to constantly re-interpret existing knowledge (Gläser, 2006: 140). For example, many changes have occurred in AMO physics in the 90s. New methods were being developed at a rapid pace, and others went out of use. One obvious major change was the experimental realization of a BEC in 1995. In the results we found solutions with clusters starting to grow enormously in size around 1995, but no ground truth could be unambiguously linked to this cluster. Hence, we cannot infer that emerging topics are reconstructed by a cluster just because it is growing over time.

Challenges for and of the algorithms

The failure of the four combinations of data models and algorithms to reconstruct our ground truths may also be caused by properties of the algorithms. Both algorithms have in common that they solely use the network topology, i.e. patterns in links (DC or BC links) between nodes (publications). Each algorithm optimizes a function used for the entire network and assigns every single node to exactly one community.[9] In other words, the algorithms are *forced* to assign publications to clusters based on the best fit regardless of how good this fit is in absolute terms. This might lead to a misassignment of publications for reasons that are purely technical, i.e. due to an 'assignment dilemma' the algorithm has to resolve.

The assignment of publications ultimately depends on the definition of a community an algorithm is based on (Schaub et al., 2017). These definitions differ between the algorithms. For the Leiden algorithm with the CPM quality function, a community is a set of nodes with a link density that is higher inside than the link density between sets. For Infomap, a community is a set of nodes where each node can be reached from the others easily with only a few steps. This seemingly minor difference may have major consequences. For example, we observed that the Infomap algorithm tends to treat nodes with a very high degree (publications with many citations or many references) as being separate from larger sets of nodes, and tends to 'let them stand alone' in their own module. Such a 'high degree' singular module serves the goal of Infomap to create the minimum description length of a random walker jumping over the modules.

Since Infomap creates a multilevel hierarchical solution, we had to decide at which level of hierarchy we evaluate the reconstruction of the ground truths. Our decision for the second-lowest level was driven by the desire to have a consistent level which creates clusters of reasonable size. Nevertheless, one could also argue to do both reconstruction experiments 1 and 2 on all levels of the hierarchy of each Infomap solution, and see on which level each ground truth is reconstructed. While we did not conduct this particular experiment, we can doubt its success because a consistent trend in the experiments we conducted is that larger clusters always lump together publications from different ground truths.

More generally speaking, it is not clear yet how these and other properties of algorithms match the properties of topics and the ways in which these properties are reflected by the data models to which the algorithms are applied. Future work should attempt to bring together insights from the sociology of science on properties of topics, knowledge about the traces topics leave in bibliometric metadata, approaches developed in network science for analysing large networks, and knowledge about the assumptions about networks on which these approaches are based.

Conclusions

Attempts to solve the task of reconstructing scientific topics from bibliometric metadata without testing approaches against a ground truth that is defined *ex ante* impairs the cumulation of knowledge on this task. In this study we provide an attempt to combine

[9] Note that there exists also a variant of Infomap that allows for overlapping communities, see Esquivel & Rosvall, (2011). Specifically, this variant allows for boundary nodes, which have been previously determined by Infomap, to be assigned to more than one community.

micro-level content-based analyses with bibliometric metadata to construct micro- and meso-level ground truths that operationalised our definition of a topic. This might represent a way forward to learn about the relation between structures reconstructed by different algorithm- and data model combinations and different topics. Unfortunately, all four combinations of data models and algorithms failed the test. Our experiments lead to the following four conclusions.

First, our attempt to construct ground truths ex ante (which, to our knowledge, is the first attempt ever) highlighted interesting methodological problems of this approach. If we want to use ground truths to validate our approaches, we must solve the problem of constructing a ground truth that represents the overlapping knowledge structures researchers share in their work. If our premise that we need to start from a firm grounding in knowledge about individual research processes is correct, the challenge is to aggregate these individual-level ground truths in order to obtain meso-level topics.

Second, the results of our macro-level experiments indicate that the clusters produced by four approaches do not represent the ground truths we constructed. The approaches created sometimes accurate and in most cases inaccurate representations of the ground truths, which indicates that some parts of some ground truths might be reconstructed. However, we do not know which ground truth is reconstructed under which conditions. This leads us to the conclusion that discussing the validity of any single approach (data model + algorithm) is likely to be fruitless because a single approach cannot be valid. Our results suggest that we need to employ several data models and algorithms simultaneously to reconstruct thematic structures from publication metadata. This combination would also lead to the reconstruction of overlapping clusters, which makes the search for a single algorithm that produces overlapping clusters (Havemann et al., 2017) obsolete.

Third, however, a reconstruction of topics from networks constructed from bibliometric metadata may not be possible at all. It could well be that the thematic structures researchers work with do not leave enough traces in the metadata to be reconstructed by bibliometrics.

Finally, we believe that in order to answer this fundamental question, the further evaluation of topic reconstruction approaches needs to proceed in parallel with enlarging the body of knowledge on properties of topics in the various scientific disciplines. At the moment, we know very little about the ways in which topics in different fields form and develop, e.g. in theoretical physics vis-à-vis experimental physics. If we had more knowledge on these properties, we could assess the bibliometric traces these kinds of topics and their developments leave.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-021-03920-3.

Acknowledgement We are grateful to Kevin Boyack and Richard Klavans for their helpful comments on an earlier version of this paper. This work was supported by the German Ministry of Education and Research (Grant 01PU17003).

Funding Open Access funding enabled and organized by Projekt DEAL.

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. *Quant Sci Stud*. https://doi.org/10.1162/qss_a_00027

Bohlin, L., Edler, D., Lancichinetti, A., & Rosvall, M. (2014). Community Detection and Visualization of Networks with the Map Equation Framework. In Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring Scholarly Impact: Methods and Practice* (S. 3–34). Springer International Publishing. https://doi.org/https://doi.org/10.1007/978-3-319-10377-8_1

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

Cherifi, H., Palla, G., Szymanski, B. K., & Lu, X. (2019). On community structure in complex networks: Challenges and opportunities. *Appl Netw Sci, 4*(1), 1–35. https://doi.org/10.1007/s41109-019-0238-9

Chubin, D. E. (1976). The conceptualization of scientific specialities. *Sociol Quarterly, 17*(4), 448–476.

Chumachenko, A. V., Kreminskyi, B. G., Mosenkis, I. L., & Yakimenko, A. I. (2020). Dynamics of topic formation and quantitative analysis of hot trends in physical science. *Scientometrics*. https://doi.org/10.1007/s11192-020-03610-6

Edge, D., & Mulkay, M. J. (1976). *Astronomy transformed: The emergence of radio astronomy in britain*. Hoboken: Wiley.

Esquivel, A. V., & Rosvall, M. (2011). Compression of flow can reveal overlapping-module organization in networks. *Phys Rev X, 1*(2), 021025. https://doi.org/10.1103/PhysRevX.1.021025

Fallani, L., & Kastberg, A. (2015). Cold atoms: A field enabled by light. *EPL (Europhys Lett), 110*(5), 53001. https://doi.org/10.1209/0295-5075/110/53001

Fleck, L. (1979). *Genesis and development of a scientific fact*. Chicago: The University of Chicago Press.

Giddens, A. (1979). *Central problems in social theory: Action, structure, and contradiction in social analysis*. University of California Press.

Glänzel, W., & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics, 37*(2), 195–221. https://doi.org/10.1007/BF02093621

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. (2006). *Wissenschaftliche Produktionsgemeinschaften*. Die soziale Ordnung der Forschung.

Gläser, J. (2020). Opening the Black Box of Expert Validation of Bibliometric Maps. *Lockdown Bibliometrics: Papers not submitted to the STI Conference 2020 in Aarhus*, 27–36. https://www.sos.tu-berlin.de/fileadmin/fg369/Jochen_Glaeser__ed__2020_Lockdown_Bibliometrics_-_Papers_not_submitted_to_the_STI_conference_2020_in_Aarhaus_SoS_Discussion_Paper_02_2020.pdf

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

Gläser, J., & Laudel, G. (2013). Life With and without coding: Two methods for early-stage data analysis in qualitative research aiming at causal explanations. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *14*(2), Article 2. https://doi.org/10.17169/fqs-14.2.1886

Gläser, J., & Laudel, G. (2015). A bibliometric reconstruction of research trails for qualitative investigations of scientific innovations. *Historical Social Research / Historische Sozialforschung Vol. 40, No. 3 (2015): Special Issue: Methods of Innovation Research: Qualitative, Quantitative and Mixed Methods Approaches*. https://doi.org/10.12759/hsr.40.2015.3.299-330

Gläser, J., & Laudel, G. (2019). The discovery of causal mechanisms: Extractive qualitative content analysis as a tool for process tracing. *Forum Qualitative Sozialforschung / Forum: Qualitative Social Research*, *20*(3), Article 3. https://doi.org/10.17169/fqs-20.3.3386

Goffman, E. (1974). *Frame analysis: An essay on the organization of experience*. Cambridge: Cambridge University Press.

Haunschild, R., Schier, H., Marx, W., & Bornmann, L. (2018). Algorithmically generated subject categories based on citation relations: An empirical micro study using papers on overall water splitting. *J Informetr, 12*(2), 436–447. https://doi.org/10.1016/j.joi.2018.03.004

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

Held, M., & Velden, T. (2019). How to interpret algorithmically constructed topical structures of research specialties? A case study comparing an internal and an external mapping of the topical structure of invasion biology. *Proceedings of the International Conference on Scientometrics and Informetrics*, 1933–1939.

Hric, D., Darst, R. K., & Fortunato, S. (2014). Community detection in networks: Structural communities versus ground truth. *Physical Review E, 90*(6), 062805. https://doi.org/10.1103/PhysRevE.90.062805

Interdonato, R., Atzmueller, M., Gaito, S., Kanawati, R., Largeron, C., & Sala, A. (2019). Feature-rich networks: Going beyond complex network topologies. *Appl Netw Sci, 4*(1), 1–13. https://doi.org/10.1007/s41109-019-0111-x

Kheirkhahzadeh, M., Lancichinetti, A., & Rosvall, M. (2016). Efficient community detection of network flows for varying Markov times and bipartite networks. *Physical Review E, 93*(3), 032309. https://doi.org/10.1103/PhysRevE.93.032309

Klavans, R., & Boyack, K. W. (2011). Using global mapping to create more accurate document-level maps of research fields. *Journal of the American Society for Information Science and Technology, 62*(1), 1–18. https://doi.org/10.1002/asi.21444

Klavans, R., & Boyack, K. W. (2017a). Research portfolio analysis and topic prominence. *J Informetr, 11*(4), 1158–1174. https://doi.org/10.1016/j.joi.2017.10.002

Klavans, R., & Boyack, K. W. (2017b). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *J Assoc Inf Sci Technol, 68*(4), 984–998. https://doi.org/10.1002/asi.23734

Koopman, R., & Wang, S. (2017). Mutual information based labelling and comparing clusters. *Scientometrics, 111*(2), 1157–1167. https://doi.org/10.1007/s11192-017-2305-2

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago: The University of Chicago Press.

Laudel, G., Lettkemann, E., Ramuz, R., Wedlin, L., & Woolley, R. (2014). Cold atoms—Hot research: High risks, high rewards in five different authority structures. In R. Whitley & J. Gläser (Eds.), *Research in the Sociology of Organizations* (Bd. 42, S. 203–234). Emerald Group Publishing Limited. https://doi.org/10.1108/S0733-558X20140000042007

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

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

Schaub, M. T., Delvenne, J.-C., Rosvall, M., & Lambiotte, R. (2017). The many facets of community detection in complex networks. *Applied Network Science, 2*(1), 4. https://doi.org/10.1007/s41109-017-0023-6

Schütz, A. (1967). *The phenomenology of the social world*. Evanston: Northwestern University Press.

Schütz, A., & Luckmann, T. (1973). *The structures of the life-world*. Evanston: Northwestern University Press.

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. https://doi.org/10.1002/asi.20994

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

Small, H. (2006). Tracking and predicting growth areas in science. *Scientometrics, 68*(3), 595–610. https://doi.org/10.1007/s11192-006-0132-y

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy, 43*(8), 1450–1467. https://doi.org/10.1016/j.respol.2014.02.005

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

Thijs, B., & Glänzel, W. (2018). The contribution of the lexical component in hybrid clustering, the case of four decades of "Scientometrics." *Scientometrics, 115*(1), 21–33. https://doi.org/10.1007/s11192-018-2659-0

Traag, V., Waltman, L., van Eck, N. J. (2018). From Louvain to Leiden: Guaranteeing well-connected communities@@@. [Physics]. http://arxiv.org/abs/1810.08473

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

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

Waltman, L., & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science: A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology, 63*(12), 2378–2392. https://doi.org/10.1002/asi.22748

Whitley, R. (1974). Cognitive and social institutionalization of scientific specialties and research areas. In R Whitley (Ed.), *Social Processes of Scientific Development* (S. 69–95). Routledge & Kegan Paul.

Whitley, R. (2000). *The intellectual and social organization of the sciences*. Clarendon Press.

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics, 63*(2), 373–401. https://doi.org/10.1007/s11192-005-0218-y