RESEARCH ARTICLE

Topics as clusters of citation links to highly cited sources: The case of research on international relations

Frank Havemann

Institut für Bibliotheks- und Informationswissenschaft, Humboldt-Universität zu Berlin, D 10099 Berlin, Dorotheenstr. 26, Germany

Keywords: citation networks, core-periphery structures, link clustering, memetic algorithm, overlapping communities, topics

ABSTRACT

Following Henry Small in his approach to cocitation analysis, highly cited sources are seen as *concept symbols* of research fronts. But instead of cocited sources, I cluster citation links, which are the thematically least heterogenous elements in bibliometric studies. To obtain clusters representing topics characterized by concepts, I restrict link clustering to citation links to highly cited sources. Clusters of citation links between papers in a political-science subfield (international relations) and 300 of their sources most cited in the period 2006–2015 are constructed by a local memetic algorithm. It finds local minima in a cost landscape corresponding to clusters, which can overlap each other pervasively. The clusters obtained are well separated from the rest of the network but can have suboptimal cohesion. Cohesive cores of topics are found by applying an algorithm that constructs core-periphery structures in link sets. In this methodological paper I discuss some initial clustering results for the second half of the 10-year period.

1. INTRODUCTION

If a topic is defined as a focus on scientific knowledge shared by a number of researchers, topics should manifest themselves in clusters of cocited sources, because cited sources represent theoretical, methodological, or empirical knowledge used or at least discussed by citing authors.

Topics can overlap in papers and even more in books if they deal with more than one topic. Another kind of overlap can occur on the level of content of topics: Shared knowledge itself can be in the foci of researchers working on different topics. We therefore need a clustering algorithm that delivers overlapping clusters.

When topics are represented as disjoint clusters of cocited sources, they overlap in papers that cite sources in different clusters. But a cited source can also correspond to more than one topic. We therefore have to allow for overlapping clusters of cited sources, which to the best of my knowledge has not been done in any cocitation analysis so far.

Cocitation analysis was independently proposed by Irina Marshakova (1973) and by Henry Small (1973). Small (1978) also introduced the notion of *concept symbols* represented by highly cited sources, for which cocitation clusters are constructed. By adding the papers that cite

an open access journal

Check for updates

Citation: Havemann, F. (2021). Topics as clusters of citation links to highly cited sources: The case of research on international relations. *Quantitative Science Studies*, *2*(1), 204–223. https://doi.org/10.1162/qss_a_00108

DOI: https://doi.org/10.1162/qss_a_00108

Supporting Information: https://doi.org/10.1162/qss_a_00108

Received: 30 July 2020
Accepted: 11 November 2020

Corresponding Author:
Frank Havemann
frank.havemann@ibi.hu-berlin.de

Handling Editor:
Ludo Waltman

The MIT Press

concept symbols in a cocitation cluster we augment the picture of the corresponding *research front* (Garfield, 1985). Cocitation analysis is the usual approach to clustering concept symbols in citation networks, but not the only possible one. I propose, instead, to cluster citation links from papers to concept symbols. Link clustering in the bipartite network of citing papers and cited sources avoids the projection onto the cocitation graph of sources and any need for normalizing and thresholding cocitation strength. From clusters of citation links between papers and sources, overlapping clusters of citing research-front papers and of cited concept symbols can be deduced. Thus, we obtain overlapping clusters of highly cited sources that are connected through papers that cocite them.[1]

Among several clustering methods that allow for overlapping clusters, link clustering has an important advantage when applied to citation networks: Citation links are the thematically least heterogenous elements in bibliometric studies. In nearly all cases, a paper cites a source due to only one knowledge claim. Even when a paper refers to two or more knowledge claims in a cited source, they often belong to one topic, especially if we search for larger and more general topics, as is done here by restricting link clustering to citation links between papers and highly cited sources.

Topic definition and the link clustering approach applied here have recently been discussed by Havemann, Gläser, and Heinz (2017). In that paper, a new evaluation function for link clusters, Ψ, and a local memetic algorithm for link clustering based on this function, PsiMinL, were proposed and tested for two kinds of citation networks: a network of direct citations in a set of astronomy papers published within 8 years, and a bipartite network of one volume of these papers and all their cited sources. I here also apply PsiMinL to a bipartite network of papers and sources, but restrict the set of sources to highly cited ones.

Clustering links in networks instead of nodes had been introduced by Evans and Lambiotte (2009) and by Ahn, Bagrow, and Lehmann (2010). In both approaches graphs are partitioned into disjoint clusters of links. From them overlapping clusters of nodes are deduced. In contrast to these global methods, PsiMinL evaluates each link cluster in a local manner independently of other clusters. It therefore can produce clusters that overlap each other pervasively (i.e., not only in their boundaries but also in inner links and nodes). A local evaluation of clusters also matches the local character of topics (Havemann et al., 2017).

Clusters or communities in networks are considered as highly cohesive subgraphs that are well separated from the rest of the network (Fortunato, 2010). There are cases where these two features of communities cannot be maximized at the same time. Methods can be classified with regard to producing well-separated or well-connected communities (Rosvall, Delvenne, et al., 2019). Like several other algorithms, PsiMinL delivers clusters that can have low cohesion (i.e., they can easily be split into two or more subclusters). This bias of the algorithm is one of the evaluation function $\Psi(L)$ for a cluster given as a link set L: It measures separation and is much less sensitive to changes in cohesion (Havemann, Gläser, & Heinz, 2019).

The evaluation function $\Psi(L)$ allows for lowly cohesive clusters, but that does not hinder its use for an evaluation of topic clusters. Not all knowledge in a shared focus has to be cited in all papers that contribute to the corresponding topic. Only those sources have to be cited that are used for the production of new knowledge. Although authors often cite other sources, too, we cannot expect that all sources in a cluster are cited in all papers contributing to the topic.

Clusters of highly cited sources that represent topics have to be well separated but can have low internal cohesion.

[1] Note that here link clustering is not applied to cocitation links between concept symbols but to citation links in the bipartite network of citing papers and cited sources.

A second argument for favoring well-separated clusters is the hierarchical structure of sets of topics. A topic can have subtopics (i.e., the splitting of its cluster should not be too difficult). Two topics can also overlap in one subtopic. Then we have no strict hierarchy but a poly-hierarchy (Havemann et al., 2017).

Nonetheless, we are interested in cohesive cores of topics corresponding to dense subgraphs of citation networks that are not necessarily well separated from the rest of the network. To extract such dense cores from a well-separated link cluster, an algorithm was proposed recently by Havemann et al. (2019). The CPLC-algorithm finds core-periphery structures of link clusters.

The analysis reported here was made within the Global Pathways project.[2] The aim of this project is to identify topic-based, language-based, and regional or national substructures in research on international relations (IR).

I must leave all conclusions regarding the content structure of IR research to a forthcoming paper enriched with the project team's IR competence (Risse, Wemheuer-Vogelaar, & Havemann, 2020). I here present the results of a test of the proposed approach. The focus of the paper is on methodological challenges.

2. DATA

For the analysis of IR literature within the Global Pathways project, we wanted to obtain a set of papers in Web of Science (WoS) that prioritizes recall over precision. The time span for all downloads was 2006–2015. We started from 115 journals indexed in the WoS category *International Relations* and added four journals from *Political Science*. In the following, these journals are referred to as *IR journals*. We also searched for book chapters in the Book Citation Index of WoS that are categorized as *International Relations*.

All documents of those types that are usually published to communicate new research results, namely articles, letters, and proceedings papers (*original papers*), were downloaded, and also reviews and book reviews. WoS also offers access to SciELO (*Scientific Electronic Library Online*, a database mainly covering publications from Latin American countries). From SciELO, records categorized as *International Relations* were also downloaded. The list of journals and further details of data can be found in the Supplementary Material.

After identifying references automatically, as described in the Supplementary Material, the 300 most highly cited sources were selected. I searched manually for further references in the data set that could be identified with them. Here references to different pages and editions of books were identified. The list of the top 300 sources can be found in the Supplementary Material. Of these, 203 have been classified as dealing with IR themes (Table 1 in the Supplementary Material). Experiments with clustering smaller numbers of concept symbols revealed that on approaching 300 highly cited sources, only peripheral topics were added and the central topic clusters had become stable.

3. METHODS

In the following I will discuss some essential elements of the two algorithms applied. Reading these sections is useful for understanding the design of the experiments and their results. Readers who are not interested in methodological details can skip this section and proceed with the results in Section 4.1. Further details can be found in the two papers mentioned (Havemann et al., 2017, 2019).

[2] http://t1p.de/globalpathways.

3.1. Link Clustering: PsiMinL Algorithm

In one sentence, PsiMinL is an evolutionary algorithm that searches in a cost landscape for local minima that correspond to well-separated link clusters. Because genetic operators (mutation, crossover, and selection) are combined with deterministic local searches in the cost landscape, PsiMinL can be called a memetic algorithm (Neri, Cotta, & Moscato, 2012). A PsiMinL glossary is in the Supplementary Material (section 2).

Each possible link set L corresponds to a place in the cost landscape. The height of place L is given by the cost function *normalized node-cut* $\Psi(L)$ of link set L. A lower value of $\Psi(L)$ signals a better separated link set L. Normalized node-cut can be defined as

$$\Psi(L) = \frac{\sigma(L)}{k_{\text{in}}(L)} + \frac{\sigma(L)}{k_{\text{in}}(E-L)}, \tag{1}$$

with

$$\sigma(L) = \sum_{i=1}^{n} \frac{k_i^{\text{in}}(L)\left(k_i - k_i^{\text{in}}(L)\right)}{k_i}, \tag{2}$$

where k_i is the degree of node i, $k_i^{\text{in}}(L)$ its internal degree with respect to link set L, and $k_{\text{in}}(L) = \sum_{i=1}^{n} k_i^{\text{in}}(L) = 2|L|$. Index i runs through all n nodes but $k_i^{\text{in}}(L) = 0$ for all nodes that are not attached to a link in L. Set E includes all m edges. Note that $\sigma(L) = \sigma(E - L)$ because $k_i^{\text{in}}(E - L) = k_i - k_i^{\text{in}}(L)$, and that $k_{\text{in}}(E - L) = 2m - k_{\text{in}}(L)$. Thus, $\Psi(L) = \Psi(E - L)$, the cost function of set L equals that of its complement $E - L$.

Deriving their link-clustering approach, Evans and Lambiotte (2009) introduced a random link-node-link walker. The first summand on the right-hand side of Eq. 1 is the probability of such a walker sitting on a link in L escaping from L and the second summand is the escape probability for the complement of L (Havemann et al., 2019). Further motivations for using the Ψ-function were given by Havemann et al. (2017).

A connected link set L that corresponds to a local minimum in the cost landscape is called a *link cluster* or a *link community*. The cost landscape is very rough (i.e., there are many local minima that differ only in a few links). We are interested in well-separated link sets that differ from any better separated set in more than only some links. Therefore we need a resolution parameter r. It is used to decide whether we can consider a link set L as a *valid* community. If there is a link set L_0 with $\Psi(L_0) < \Psi(L)$ and the two link sets differ in less than $r|L|$ links then L_0 makes L *invalid*. In other words, we search for local minima with no lower place in the landscape within a radius $r|L|$.

A local search in PsiMinL is done by greedily including neighboring links to a connected link set L or by excluding links from L that are attached to boundary nodes. Here I have implemented a procedure that tries to lower cost in an alternating sequence of link exclusion and inclusion until no further improvement is possible.

A simple local search—done by going downhill in the cost landscape—is soon trapped in the next local minimum. We allow the greedy algorithm in local searches to proceed even when the costs are rising. It stops and goes back to the place L_{min} of the last cost minimum in the search if it does not find a place with lower cost after $r|L_{\text{min}}|$ steps (i.e., if it does not find a link set that makes L_{min} invalid). In other words, the local search can tunnel through barriers in the cost landscape if the end of the tunnel is not too far away. Then the link set at the end of the tunnel invalidates the cluster at the tunnel entry.[3]

[3] In Figure 4 in the Supplementary Material a cost-size diagram of a local search visualizes how the sequence of greedy exclusion and inclusion of links proceeds and how the search path tunnels through barriers in the cost landscape.

In memetic algorithms, deterministic local searches are combined with evolutionary genetic operators (i.e., with mutation, crossover, and selection). We need randomness because even tunneling does not avoid trapping of local searches in local minima corresponding to invalid communities. A population is initialized from a seed subgraph by a local search followed by mutations and again local searches until the desired number of different individuals is reached. Mutation and crossover are used to explore the cost landscape around a preliminarily valid cluster at a local minimum that corresponds to the current best individual of a population.

If two clusters have well-separated boundaries, their intersection and their union could also have such a boundary. Therefore, offspring are made from the intersection and union of parents. As one parent the current best individual is chosen, while the other is selected from among those individuals that have large genetic distance (measured as set difference) from the best individual. After mutations and crossovers (both followed by local searches) the best individuals are selected for the next generation.

The memetic algorithm PsiMinL was implemented as an R-package[4] with parallel procedures for all members of a population that undergoes an evolution. Because each cluster is evaluated independently from all others, several evolutions starting from different seed subgraphs can run parallel too. As seed subgraphs one can use clusters obtained from any fast clustering algorithm. The set of all valid clusters is totally independent of the set of seeds used to find them, but there is no guarantee of finding all valid clusters with a given seed set.

Different runs of PsiMinL starting from the same seed can end in different local minima of the cost landscape. Tests of PsiMinL on the cost landscape of a large citation network of 8 years of astronomy papers (Havemann et al., 2017) show two typical cases of path bifurcation. The algorithm can run into different hollows, or it may end at different places in the same hollow. In the second case, the distances between different minima were found to be small, often much smaller than the resolution radius $r|L|$. This means that we can assume that further runs of PsiMinL will improve and change a result only slightly.

To maintain an overview over the many experiments necessary for finding as many valid clusters as possible in a network, it is convenient to ensure that in a local search starting from a mutant or from an offspring of the current best cluster L_0 and ending in a better one, L_0 is invalidated. Consequently, if the first place on the path downhill with a cost $\Psi < \Psi(L_0)$ is not within a radius $r|L_0|$, then the local search is stopped and the individual link set is excluded from further evolution.

PsiMinL has many parameters (population size, mutation variances and rates, number of crossovers, etc.) but only resolution r influences the results. All other parameters only influence the time needed to obtain them.[5]

Recently, Gabardo, Berretta, and Moscato (2020) have proposed a new memetic algorithm for global link clustering resulting in overlapping communities of nodes. They evaluate whole disjoint link partitions with the density metric proposed by Ahn et al. (2010). Chalupa, Hawick, and Walker (2018) have tested different crossover operators combined with deterministic and randomized variants of local search for finding bottlenecks in networks that correspond to minima of conductance Φ, an evaluation function that favors well-separated subgraphs in the world of node clustering as normalized node-cut Ψ does for link clustering. They found "sparse

[4] The yet unpublished R-package PsiMinL (programmed by Andreas Prescher) and a detailed description of it and its installation will be delivered on request.

[5] Table 6 in the Supplementary Material lists parameters, their meanings, and their values chosen in the experiments described below.

imbalanced cuts into a community and the rest of the network, as well as relatively balanced partitions" (p. 28 in preprint version). Like that of Lu, Hao, and Wu (2020) but in contrast to PsiMinL, their algorithm randomly selects genes of parents for offspring clusters and applies mutation only for population initialization. Further papers related to the algorithm PsiMinL are referred to by Havemann et al. (2017, p. 1095). Evolutionary algorithms used for detecting communities in networks have been reviewed by Clara Pizzuti (2017).

Like conductance Φ, normalized node-cut Ψ neglects the direction of links. Thus, applying it to a bipartite network of papers and their cited sources means that papers and sources are treated symmetrically.

3.2. Cores and Peripheries of Link Clusters: The CPLC Algorithm

CPLC constructs core-periphery structures (named *towns*, for short) in a given link set as nested subgraphs with decreasing cohesion. Large star subgraphs have a high local density of links. This density notion is the translation of usual graph density into the world of link clustering (Havemann et al., 2019, p. 5). For a recent review of algorithms for core-periphery construction see the paper by Tang, Zhao, et al. (2019).

In our case the largest stars are highly cited sources with their incoming citation links. A town is defined as a size-ordered cluster of stars where two stars are never indirectly connected via smaller stars only. To illustrate this definition, we can imagine the size of stars as the height of hills. Then all smaller stars of a town can be reached from the largest one on a path that is never going uphill.

A star is connected to a town if it shares a minimum number of outer nodes with the set of town stars of equal or larger size; otherwise it becomes the center of an independent town. The minimum number of outer nodes is determined by a resolution parameter q with $0 \leq q < 1$, which is used as the minimum threshold of relative overlap for a star to be attached to a town.

Instead of arbitrarily setting parameter q, its whole range is explored by starting with minimal resolution $q = 0$ and increasing it recursively to a value at which it is possible to obtain at least one more town in the given link set. To choose a resolution level at which useful core-periphery structures are constructed, different criteria can be applied. One can, for example, consider towns at a level where the two largest stars in the link set are centers of different towns.

Towns of clusters can also be used to construct appropriate small seed subgraphs for PsiMinL.

4. EXPERIMENTS

4.1. Link Clustering

I divided the period 2006–2015 into two 5-year periods for two reasons. First, because 5 years is long enough to diminish the influence of random fluctuations of citation data. Second, because a comparison of the two 5-year periods can be made.[6]

Any paper that cites only one of the top 300 sources can be neglected when clusters of them are constructed. For clustering citation links to these sources, PsiMinL only needs papers that cite at least two of them. For 2006–2010 there are 4,778 such papers and 6,494 papers for the last 5 years. Only papers in IR journals and books were included.

[6] In addition, IR experts can better compare clusters obtained for this period with the results obtained by Kristensen (2018) who analyzed author cocitation in IR-papers published in 2011–2015.

Seed subgraphs were made from disjoint clusters of cited sources that have been obtained by applying Ward clustering to the cocitation network of top 300 sources. Distances were calculated from the *similarity of views* (Gläser, Heinz, & Havemann, 2015).[7]

Usually, an optimal cut through the whole dendrogram of a hierarchical clustering is chosen to get a partition of a network. I have tested this approach to seed construction, but starting from 15 middle-sized seeds, most evolutions had a long path through the cost landscape: The resulting clusters have sizes very different from their seeds (cf. Supplementary Material, Figure 7). Clusters of one cut through the dendrogram are not well suited as seed subgraphs for an algorithm that results in a poly-hierarchy of clusters. Therefore I have applied an alternative method: For different numbers of clustered top 300 sources, the Ward clusters with the longest branches in the dendrogram were selected for constructing seed subgraphs for link clustering. A Ward cluster has a long branch if it has relatively low variance and if the next larger cluster in the hierarchy has clearly larger variance. Low variance means strong cohesion, while large variance of the next supercluster means weak cohesion or the chance that its subclusters are well separated from each other.[8] The selection of 27 Ward clusters for seed construction is described in the Supplementary Material (section 3).[9]

For any selected Ward cluster of cocited sources the set of citation links to all its sources was used as a seed subgraph for link clustering. PsiMinL first makes a deterministic local search starting from a seed, and then makes an evolutionary search. For a second run of memetic search I made additional seed subgraphs from intersections and unions of valid clusters. I also used selected core-periphery structures constructed by applying CPLC on valid clusters as seed subgraphs.

In all previous experiments, we had fixed the resolution parameter on one level: $r = {}^1/_3$. Here I allowed for several levels of resolution. First, resolution parameter $r = {}^1/_{20}$ was chosen, which separates all clusters that differ in at least ${}^1/_{20}$ of their links. For each seed, 16 independent evolutions were started with populations of eight individual connected subgraphs given as link sets. An evolution was stopped when during 100 generations the best individual could not be improved. In the next phase, the eight best of 16 resulting individuals formed a new population. This was repeated until most of the 16 evolutions gave the same result.[10]

Then, the whole procedure was repeated but now with a larger resolution parameter r and using the results of the first run as seeds. I made such iterations on resolution levels with $r = {}^1/_{10}$, ${}^1/_5$, ${}^1/_4$, and ${}^1/_3$. At each step of iteration a stronger condition for validity was applied than in the preceding step. All valid link clusters for, for example, $r = {}^1/_4$ are also valid for $r = {}^1/_5$, but not the other way round.

The workflow of the whole procedure including pre- and postprocessing is visualized in Figure 1. The details of the PsiMinL algorithm have been notated as pseudocode by Havemann et al. (2017, p. 1094). To give an impression of memetic evolution, the search path starting from a large seed is described and visualized in the Supplementary Material (section 4).

Figure 2 shows the costs Ψ and sizes of all 27 selected Ward seed-subgraphs, of results of initial local searches and of memetic searches on intermediary resolution levels, and of 11

[7] Ward clustering of views was done by Michael Heinz. Its results can be downloaded as R-object cv.RObj from https://zenodo.org/record/4181930 (Havemann, 2020).

[8] Branch length measures cluster quality (Havemann, Gläser, et al., 2012, p. 8).

[9] In addition, the set of seeds has been extended by including a further 23 Ward clusters with shorter branches in the dendrogram. The results are in the Supplementary Material (section 3).

[10] The parameters used are listed in Table 6 in the Supplementary Material). They had been proven as suitable in a series of previous experiments, but until now no systematic exploration of the parameter space of PsiMinL has been made.

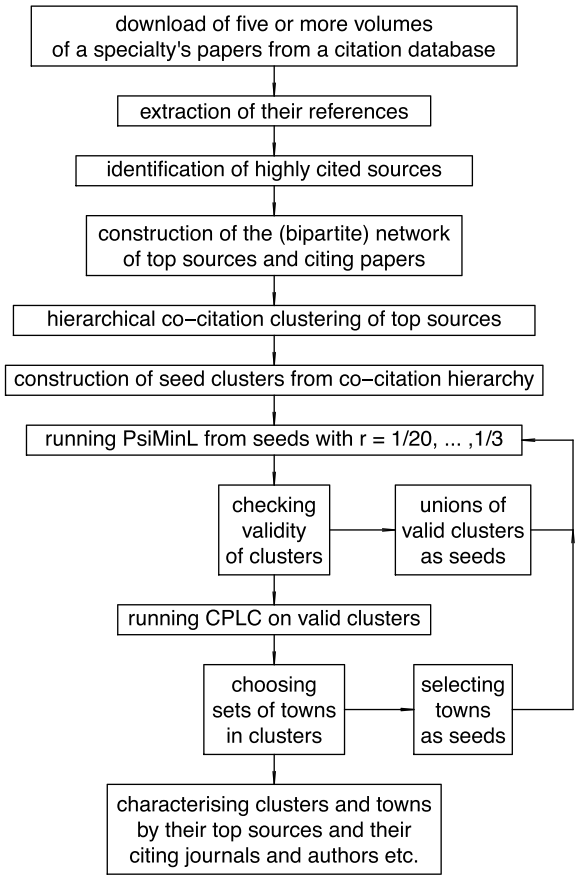

Figure 1. Workflow.

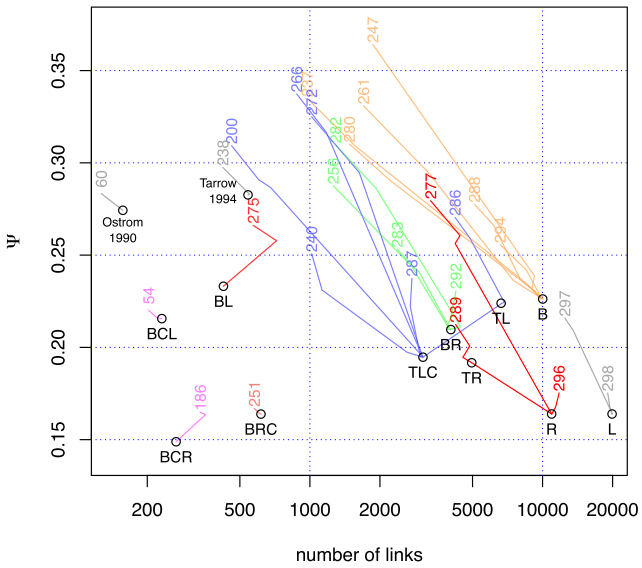

Figure 2. Cost-size diagram of 27 seed subgraphs and of steps towards 11 clusters valid at all resolution levels $r \leq {}^{1}/_{3}$.

resulting clusters on final resolution level ($r = {}^1/_3$). Each seed is connected by a line with its intermediary results and its final cluster. The colors of lines are equal for all evolutions with the same final cluster. Cluster **L** (the largest one), for example, is reached by starting memetic evolution from two large seeds with identifiers 297 and 298 (cf. Figure 3 in the Supplementary Material). Seed 298 is the largest seed (175 of the top 300 sources) and includes seed 297 (103 sources; see Figure 2 in the Supplementary Material). There are 171 sources with more than 95% of their citation links in **L**, and 160 of them are also in seed 298 (91% of 175).

Clusters **TL** and **TR** are not valid at the final resolution level but for $r = {}^1/_{10}$ and $r = {}^1/_4$, respectively. For the next levels, PsiMinL found a path through the cost landscape that ends in clusters **TLC** and **R**, respectively. All other clusters at intermediary levels are not considered here. They do not differ much from the final clusters or are valid for $r = {}^1/_{20}$ only.

The first part of Table 1 lists data for all 13 clusters that have been reached from any of the 27 selected seeds. For clusters reached from more than one seed, the first column gives the id number of the seed that is nearest in size to the final cluster. In some cases, different evolutions ended up in slightly different variants of a cluster. The best one invalidates the other variants.

According to the definition in Eq. 1 the cost function is equal for a link set and its complement. Therefore, each complement of a cluster is also a cluster if it is a connected subgraph. Indeed, the largest valid cluster (with more than half of all links) is the complement of the second largest one: $\mathbf{L} = E - \mathbf{R}$.

Complements of small subgraphs are nearly as large as the whole network and therefore not really interpretable as topics. We therefore only consider the complement of cluster **B** (on size rank 3, with about one third of all $m = |E| = 30{,}835$ citation links). $E - \mathbf{B}$ is connected, but we have to test whether it survives a local and a memetic search. That means, we have to use it as a seed subgraph for PsiMinL. $E - \mathbf{B}$ remained unchanged and therefore valid till resolution level $r = {}^1/_5$. At level $r = {}^1/_4$, PsiMinL invalidated $E - \mathbf{B}$: It found a never rising path (with tunnels) through the cost landscape ending in **R**.

The bipartite network of papers and sources is very large. Therefore, clusters are visualized on a projection of the bipartite network onto the cocitation graph of top 300 sources (Figure 3). This has the wanted side-effect that a visual comparison of the two approaches can be made (see also footnote 1). We expect link-cluster boundaries to prefer regions of sparse cocitation relations. Following Marshakova (1973), edges between the 300 selected sources were weighted with their cocitation numbers diminished by expectation values derived from a null model of independent citations. Only edges that are significant at a 95% level have been used as input for the force directed placement of nodes.[11]

The red line in the graphs of Figure 3 marks the boundary between **R** on the right-hand side and its complement **L** on the left-hand side of the graph. It connects 22 bridging sources that are cited by papers on both sides, beginning with Kant and von Clausewitz at the top and ending with Vachudova. More specifically, each of the bridging sources has not less than 5% of its citation links in each of the two complementary link clusters. All other sources have more than 95% in **L** or in **R**, respectively.

Source labels are displayed for the centers of 31 core-periphery structures obtained by running CPLC on the whole bipartite network and using results from the resolution level where the two most cited sources (Waltz 1979, Wendt 1999) become independent of each other. I have added labels for three sources at the ends of the red line (mentioned above) and for three sources

[11] Fruchterman-Reingold algorithm, implemented in R-package *sna* (Butts, 2016).

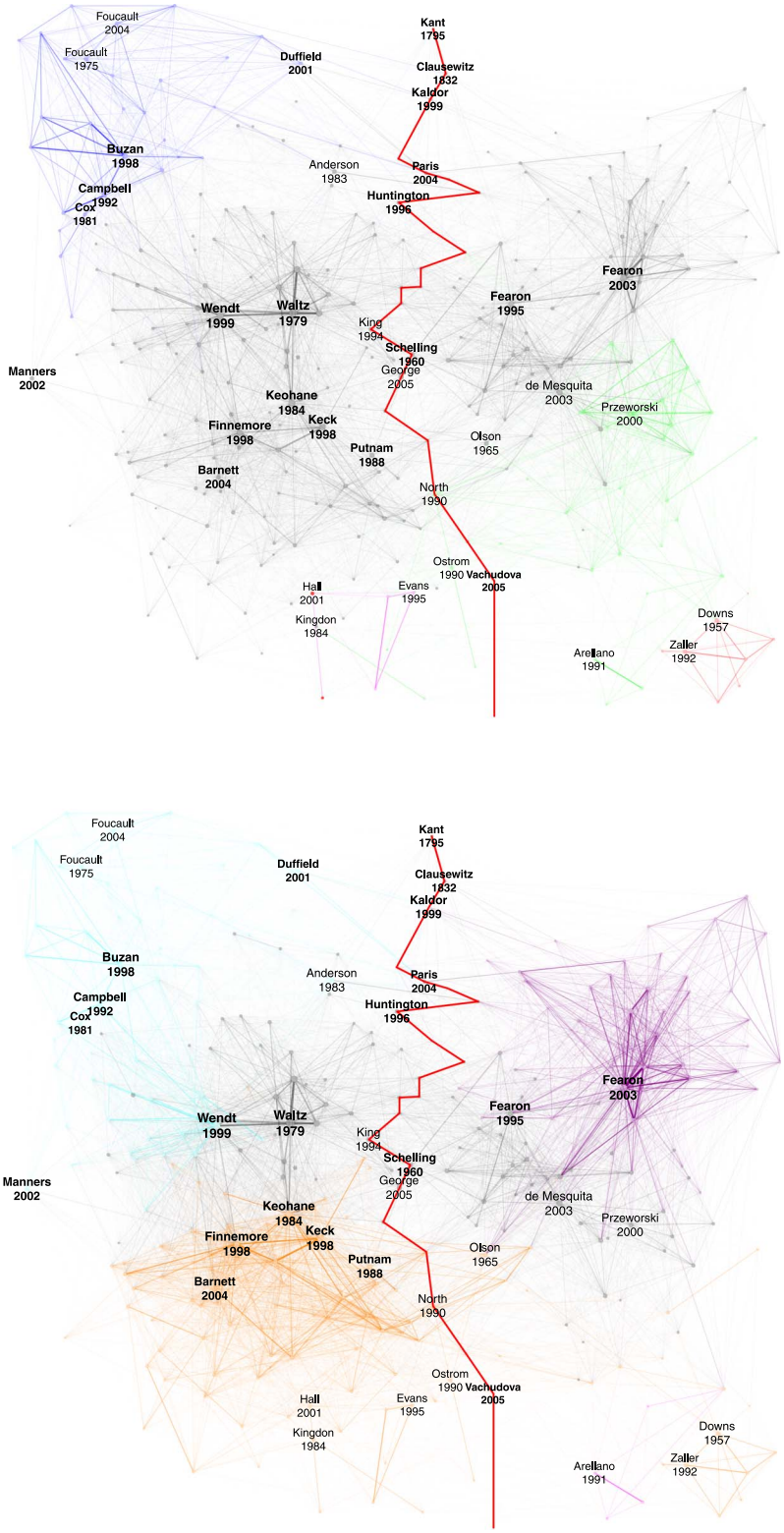

Figure 3. Graphs visualizing valid clusters 2011–2015. The red line marks the boundary between largest cluster **L** on the left and its complement, the second largest cluster **R** on the right. On the upper graph clusters **TLC**, **BR**, **BL**, **BCL**, and **BRC** are also visualized; on the lower graph clusters **TL**, **TR**, **B**, and **BCR** are visualized (cf. text).

Table 1. Link clusters ordered by seeds: first part—Ward clusters of views; second part—unions of valid clusters; third part—CPLC-towns (cf. text)

Seed	Name of cluster	Number of links	Ψ
54	**BCL**	231	0.21559
60	Ostrom 1990	157	0.27430
186	**BCR**	266	0.14886
238	Tarrow 1994	542	0.28265
251	**BRC**	616	0.16387
275	**BL**	425	0.23313
286	**TL**	6,634	0.22394
287	**TLC**	3,062	0.19469
289	**TR**	4,961	0.19167
292	**BR**	4,034	0.20966
294	**B**	10,015	0.22622
296	**R**	10,940	0.16392
298	**L**	19,895	0.16392
BCL ∪ **BCR**	**BC**	501	0.17894
BCR ∪ **BRC**	**BRB**	893	0.15913
TLC ∪ **TR**	**T**	8,027	0.20321
	Cox 1981	536	0.27489
	North 1990	265	0.33104
	Olson 1965	304	0.35011

that are centers in clusters (Arellano 1991, Evans 1995, Przeworski 2000). Labels are highlighted in bold for cited sources classified as belonging to the IR specialty.

A cluster is marked by coloring sources that have more than 95% of their citation links inside its link set. A cocitation edge is colored if more than half of all its cociting papers have citation links (to the two sources) that belong to the cluster's link set. The color used in Figure 2 for a cluster is the same as in the graphs.[12]

Bold cluster names are derived from the position in the graphs of Figure 3: **L**—left, **R**—right, **B**—bottom (orange), **TR**—top right (violet), **TL**—top left (turquoise), **TLC**—top left corner (blue). In the upper graph in Figure 3, the red links and nodes represent cluster **BRC** (bottom right corner).

[12] Citation numbers of the set of 300 selected sources restricted to citation links in valid clusters can be downloaded as R-object ccs-v7.RObj from https://zenodo.org/record/4181930 (Havemann, 2020). The file read-me.R contains R-code for listing core sources of clusters. The data set on Zenodo also includes lists of sources in clusters and on their boundaries (file Havemann2020topics.pdf) and lists of journals with numbers of papers citing sources in clusters (file citing.journals.of.clusters.pdf).

Pink elements correspond to cluster **BCL** (bottom center left). Cluster **BCL** is also a subgraph of cluster **BL** (bottom left, pink and dark red). All these small clusters are subgraphs of **BR**, which therefore is visualized not only by green nodes and cocitation links but includes all colored elements in the bottom right of this graph.

There are two small clusters in the first part of Table 1 that are named after their most cited source, both with relatively high Ψ-values:

Cluster "Tarrow 1994" includes Sidney G. Tarrow's book *Power in movement: Social movements and contentious politics* and five other sources with related themes, all outside IR and inside cluster **TR**.

The cluster "Ostrom 1990" contains two sources with all their citation links: Elinor Ostrom's famous book *Governing the commons* (90 citations) is cocited in 21 papers with *The tragedy of the commons*, the paper by Hardin Garrett published in 1968 in *Science* (37 citations). Twenty-two other sources have citation links within this cluster but get fewer than five citations from 106 papers belonging to it. The node with label "Ostrom 1990" can be found in the upper graph of Figure 3 near cluster **BL** (bottom left, pink and dark red).

The second part of Table 1 lists data of new clusters reached by starting PsiMinL from seeds that are unions of valid clusters in the first part. Unconnected unions cannot be seeds.

The three smallest clusters and **BRC** do not overlap each other in citation links, but one methodological book (Wooldridge 2002) is cited in all four clusters (by 30 papers in **BCR**, by one paper in each of the other three clusters). Thus, any union of them is a connected subgraph and can be used as a seed.

Seeds made from unions of cluster "Ostrom 1990" with each of the other three small clusters did not bring any new result. In all three cases, cluster "Ostrom 1990" was excluded already at the first resolution level ($r = {}^{1}/_{20}$) and the other cluster was reached again by memetic search.

The union of **BCL** and **BCR** has 497 citation links and $\Psi(\mathbf{BCL} \cup \mathbf{BCR}) \approx 0.18079$. PsiMinL found the slightly better cluster **BC** (bottom center) on a short path through the cost landscape and already at resolution level $r = {}^{1}/_{20}$. All these statements hold analogously for cluster **BRB** (bottom right) which is not far from the union of **BCR** and **BRC**. Starting PsiMinL from **BCL** ∪ **BRC** ended up in cluster **BRC** itself already at the first resolution level.

Both new clusters, **BC** and **BRB**, do not differ much from their seeds, which are (connected) unions of disjoint link sets. Thus, we can assume that they can easily be split into well-separated parts. Indeed, running CPLC on, for example, **BC** results in two towns very similar to **BCL** and **BCR**, respectively, already on resolution level $q = 0$. Therefore, we can expect that clusters **BC** and **BRB** are thematically not very homogeneous. This can also be said about a cluster obtained from the union of cluster "Tarrow 1994" with cluster **BCL** (678 links, $\Psi \approx 0.25973$).

Clusters **TLC** and **TR** overlap in only 24 links. Their union used as seed resulted in a new cluster with 8,027 links (**T**, for top), which is valid on all levels. Some 96% of all links in **TLC** and 89% of all links in **TR** are also in **T**.

I conclude that seeds that are connected unions of disjoint (or nearly disjoint) link sets are not useful for identifying homogeneous topics.

Other (nontrivial) unions of overlapping clusters did not result in any new valid cluster. The same holds for intersections of valid clusters. Starting PsiMinL from intersection **BR** ∪ **L**, for example, ended up with **BC**. I did not consider intersections of valid clusters that contain only a few links or more than 70% of the links of the smaller cluster because one can then expect that PsiMinL only finds this smaller cluster again.

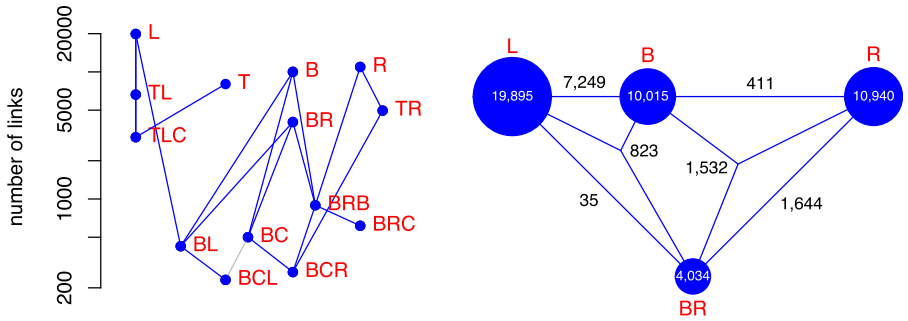

Figure 4. Left: poly-hierarchy of clusters (note log-scale of size); right: overlaps of four clusters without superclusters (size of triple overlap subtracted from size of pairwise overlaps, cf. Table 2).

The left-hand side of Figure 4 visualizes the poly-hierarchy of clusters. A blue line is drawn if the smaller cluster has less than 5% of its links outside the larger cluster.

The tiny cluster **BCL** has 215 of its 231 links (93.1%) in **BC** and is totally included in **BL**. Total inclusion is the exception. This is due to the normalization in Eq. 1. The cost of the smaller cluster is lower with some additional links, but not the cost of the larger cluster, because a smaller link set has a larger relative increase of the denominator $k_{\text{in}}(L)$ by including links than a larger set.

On the right-hand side of Figure 4, overlaps between four clusters are displayed that are not (nearly totally) included in a larger cluster. **L** and **R** have zero overlap by definition. Eight hundred and twenty-three of all 858 links in **L** ∪ **BR** are also in **B**. The remaining 35 citation links are visualized by the direct edge between **L** and **BR**. The edge betwen **B** and **BR** is missing because all 2,345 links in **B** ∪ **BR** are either in **L** or in its complement **R** (see Table 2).

4.2. Core-Periphery Structures

Constructing the core-periphery structures of a cluster can reveal its highly cohesive cores if it has one or more such cores. Clusters in the second part of Table 1 decay into two well-separated subclusters. We can therefore neglect them when we look for cohesive cores.

For all other 11 valid clusters found at resolution level $r = {}^1/_3$, core-periphery structures (towns) were constructed by running CPLC for a sequence of values of resolution parameter $q \in [0, {}^1/_2]$. Figure 5 shows the four towns in **TLC** obtained by CPLC at resolution level $q = 0.183$. The pale blue town around Foucault (1975) has a larger periphery than the three other

Table 2. Overlaps of four valid clusters without superclusters (cf. text and Figure 4)

Link set	Links
L ∪ **B** ∪ **BR**	823
R ∪ **B** ∪ **BR**	1,532
L ∪ **B**	8,072
L ∪ **BR**	858
R ∪ **B**	1,943
R ∪ **BR**	3,176
B ∪ **BR**	2,355

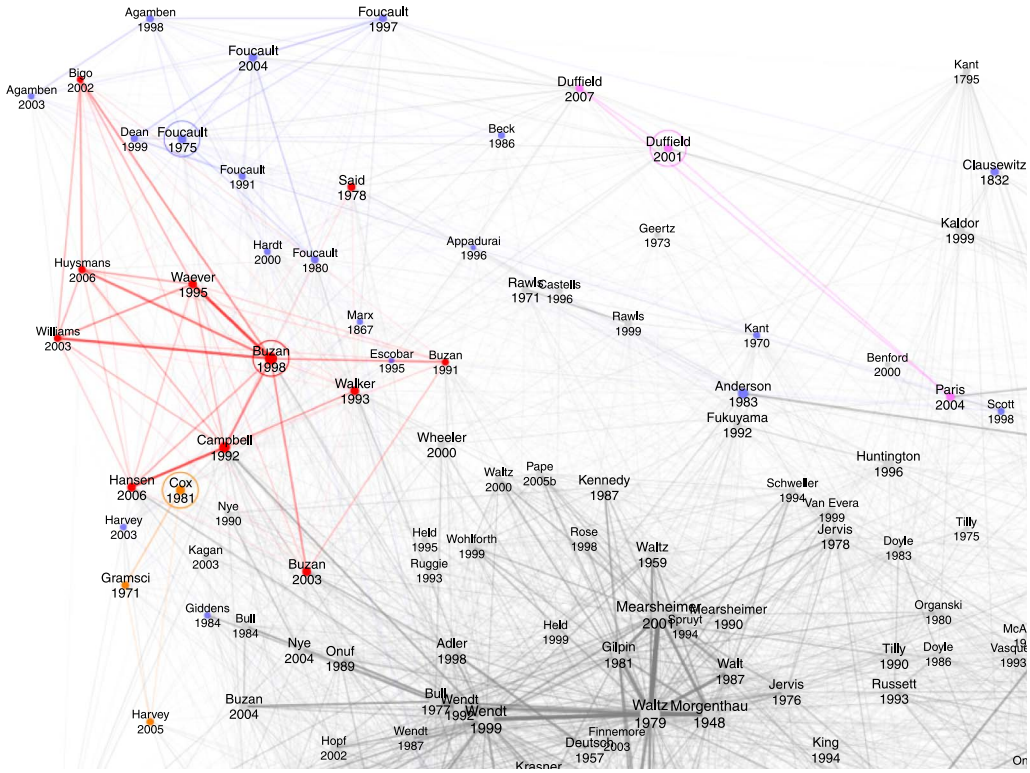

Figure 5. Four core-periphery structures (towns) in cluster **TLC** visualized by different colors. Central sources are marked by a circle.

towns. I here present only this example, which at least gives cursory evidence that CPLC indeed reveals core-periphery structures in clusters. I leave a detailed examination of results to further work.

Towns of clusters were also used as seed subgraphs for finding further clusters. One example is a town of **L** with Wendt (1995) as the center. Starting from this seed, PsiMinL rediscovered cluster **TL**. I selected those towns as seeds that promised to lead to new clusters from inspecting the cocitation graph (Figure 3). Further successful cases are the three clusters in the third part of Table 1, which are named after the centers of their seed towns.

The paper by Robert Cox (1981) about *Social forces, states and world orders* can be found on the left-hand side of graphs in Figures 3 and 5. It is often cocited with Marx and Gramsci and with two books by David Harvey published in 2003 and 2005, respectively. These five sources are the sources with full membership in this cluster and also with all their citation links inside cluster **TLC**.

The book by Douglass North (1990, on the red line in Figure 3) is significantly often cocited with the book by Oliver Williamson (1985), both dealing with economic institutions. They have all their citation links in this cluster. The next relevant source is Ostrom's book (1990), which is cited by 10 cluster papers but gets 90 citations in the whole set.

Mancur Olson's book about *The logic of collective action* (1965, on the right-hand side of the red line in graphs of Figure 3) is the only full-member source in its cluster. In contrast to the other two clusters in the third part of Table 1, this cluster remains valid only till $r = {}^1/_5$. For $r = {}^1/_4$, PsMinL invalidated it by reaching **BCR**.

5. DISCUSSION

5.1. Clustering Method

Methods for the clustering of networks can use global evaluation functions that evaluate whole partitions, such as modularity, or local functions that evaluate each cluster independently from others, such as conductance or normalized cut for node clustering (Fortunato, 2010) and normalized node-cut Ψ for link clustering.

Topics are locally defined. This favors the use of local evaluation functions for topic reconstruction. Citation links are the thematically least heterogeneous bibliometric elements. This suggests applying link clustering algorithms in citation networks. Topics can overlap and form a poly-hierarchy, which in turn means that topic clusters should not be too hard to split into subclusters. Thus cohesion cannot be the main criterion for evaluating a cluster. To date, PsiMinL is the only algorithm that is in line with all these demands. The price paid for this is long running times, the need for many CPUs, and a high complexity of the whole analysis (see also the discussion of computer running times of PsiMinL in the Supplementary Material, section 6).

Next to these abstract and technical considerations, the crucial test relates to domain knowledge: Can experts interpret not only single clusters but also the poly-hierarchy they form and their overlaps?[13] I leave this for further work.

This paper makes several novel contributions. For the first time, I apply PsiMinL to a bipartite network of highly cited sources and papers citing at least two of them. I argue that this restriction is possible because top-cited sources serve as symbols for shared knowledge of a scientific community in a field and shared knowledge is what a topic defines. This restriction reduces the network size (by a factor of 10) and therefore also the computational effort. Also, for the first time, I overcome the somewhat arbitrary choice of a fixed resolution by going through a sequence of resolution levels and using the resulting clusters on one level as seeds for the next one. A further novelty is that I construct initial seed subgraphs from clusters corresponding to long branches in the dendrogram obtained by Ward cocitation clustering. This is also the first PsiMinL analysis of a specialty belonging to the social sciences.

5.2. Clustering Results

Three different data models were used here, namely

- the bipartite network of top 300 sources and all papers in IR journals and books citing at least two of them (used by link clustering algorithm PsiMinL, leading to a poly-hierarchy of clusters);
- the projection of the bipartite network onto the cocitation graph of top 300 sources (on which clusters are displayed after selecting significant links); and
- a distance matrix between top 300 sources made from the cocitation projection weighted with Salton's cosine (used for constructing seed subgraphs from Ward clusters of views).

In spite of data differences, each link cluster concentrates in a certain region of the cocitation graph. Most clusters have boundaries going to sparse regions of the graph. This is a first hint that PsiMinL applied on a bipartite network of papers and top-cited sources leads to reasonable

[13] Otherwise, all the effort becomes problematic. A further interesting question is whether one finds top sources in overlaps that are cited for different reasons in different overlapping clusters, which was one of our arguments for clustering citation links.

clusters. I leave any further evaluation of contents of PsiMinL clusters and of their core-periphery structures obtained here to IR experts (Risse et al., 2020).

I can, however, compare these clusters quantitatively with all clusters of views on all levels of hierarchical Ward clustering. How many top 300 sources of a Ward cluster are core members of any link cluster? The results are presented in the Supplementary Material (section 7).

Three link clusters are never the best match of a seed, namely those made from unions of two clusters: **BC**, **BRB**, and **T** (second part of Table 1). This corresponds to their probable thematic inhomogeneity discussed above.

There are five exact matches between clusters, which all have fewer than seven cited sources (Table 4, section 7 of Supplementary Material). The worst match is with cluster **TL** (Salton's cosine $s \approx 0.76$). The division between the two largest clusters **L** and **R** is matched with values of $s > 0.9$.

All but one of the matched link clusters in the first part of Table 1 are matched best by their (nearest) seed. Only **TLC** is best matched by a Ward cluster that is not in the set of 27 long-branch seeds but among the 23 seeds with shorter branches (cf. footnote 9). PsiMinL reaches **TLC** from this seed too.

How can we interpret these good matches between link clusters and some Ward clusters of views that correspond to long branches in the dendrogram?

First, the two approaches are compatible and therefore supporting one another.

Second, the use of long-branch clusters as seed subgraphs for PsiMinL is confirmed as an efficient method. Starting from seeds from a global cut through the dendrogram needs longer paths in the cost landscape and resulted only in a subset of valid link clusters obtained with long-branch seeds. That means that by starting from long-branch seeds we rediscover all clusters that were found with global-cut seeds. In other words, similarity of seeds and resulting clusters is not the reason for finding this set of clusters.

Experiments with seeds corresponding to 23 branches with submaximal length in their size classes showed that we can find more small valid link clusters when starting from small seeds with shorter branches too (cf. Supplementary Material, section 3). Some of these small clusters are not as well separated as the best clusters in Table 1. Their Ψ-values exceed $^1/_4$ (cf. also Table 2 in section 3 of Supplementary Material).

The evaluation function Ψ is always larger than the escape probability of the random link-node-link walker (Evans & Lambiotte, 2009), and for small clusters only slightly larger, because the denominator of the second term in the definition of Ψ (Eq. 1) is very large. That means that for $\Psi < {}^1/_2$ the random walker's probability of remaining within the cluster is always larger than of escaping from it at the next step ($P_{\text{esc}} < {}^1/_2$).

An ordinary random walker hopping from node to node escapes from a *weak* node community as defined by Radicchi, Castellano, et al. (2004) also with a probability $P_{\text{esc}} < {}^1/_2$. Translating the definition of weak communities into the language of link clustering (Havemann et al., 2019), we can deduce that all clusters obtained here are link communities in the weak sense.

Recently Kristensen (2018) determined disjoint cocitation clusters of 332 authors highly cited as first authors in 106 IR journals in the period 2011–2015. His aim was to visualize the "communicative-sociological structures" of the discipline. He admits that neglecting coauthors of highly cited first authors can cause biases towards some authors, especially towards authors of theorizing works. He found some authors with a "fairly stable position in

the network" but others "whose work is used for positioning by several camps may shift camps depending on the specific threshold values" (p. 247).

In my approach each highly cited work can appear in more than one cluster because I produce overlapping clusters of cited sources. Topics overlap in authors even more than in papers or books, but at first glance both networks show at least some similar structures. The contents of Kristensen's *camps* of authors and of link clusters obtained here cannot be compared without knowledge of the field.

6. CONCLUSIONS

Can PsiMinL be recommended for finding a poly-hierarchy of overlapping research topics of a specialty? The experiments made in this study suggest that we indeed obtain reasonable results by applying PsiMinL to a bipartite network of selected concept symbols and all papers citing at least two of them.[14] IR experts were able to interpret them (Risse et al., 2020). All resulting clusters were only slightly changed after adding missing links to the network (see Supplementary Material, section 6). Several link clusters have a good match with Ward clusters of views (see Supplementary Material, section 7). A comparison with results of further clustering algorithms applied to the same data would be useful for evaluating the new approach to clustering concept symbols. A first trial with classic cocitation analysis (single linkage of cosine weighted links) as done by Small and Sweeney (1985) was made. Also here, the results suffer from chaining, the well-known disadvantage of single linkage. Differences between clusters obtained by PsiMinL and by other algorithms could be evaluated by experts of the specialty. I leave such comparisons to further work.

Generally, any partition of a network into disjoint clusters cannot be compared as a whole with a poly-hierarchy of overlapping clusters. A good matching of all clusters is only possible if the clusters used for a quantitative comparison form a hierarchy that has many levels (like the Ward cluster of views discussed above).

Similar results of different clustering methods can be seen as offering mutual support, but different results do not falsify any of the methods. They can be interpreted as reconstructing legitimate alternative perspectives on the structure of a specialty's literature (Gläser, Glänzel, & Scharnhorst, 2017). At most, one method could be judged as more accurate than the other when we compare both with regard to the purpose of clustering (Waltman, Boyack, et al., 2020). A poly-hierarchy of independently evaluated clusters, as delivered by PsiMinL, could represent already different perspectives on the analyzed literature.

The evaluation function Ψ can be justified within the model of a random walker who should leave a cluster with low probability (Havemann et al., 2019). To find node clusters, each step of a random walker starts and ends on a node. Link clusters can be constructed by starting and ending on links (Evans & Lambiotte, 2009).[15] Random walks are long in well-separated clusters. When a cluster contains subclusters that are only weakly connected with one another the chance of leaving it can nonetheless be as low as of leaving any of the two subclusters. In this sense random walkers are insensitive to the inner cohesion of clusters. I argue that we need cohesion

[14] One caveat has to be made: Researchers in IR, as in other specialties in social sciences, often refer to books as concept symbols. Some 175 of the top 300 sources are books (see Table 1 in the Supplementary Material). Thus, the success of the approach for specialties of natural science can be expected but not guaranteed.

[15] Recently, a random link-node-link walker's escape probability was used by Enders, Havemann, et al. (2020) to cluster 39 standard hypotheses about biological invasions for mapping this specialty.

insensitivity if we want to obtain hierarchically organized sets of clusters. Only the smallest clusters can be expected not to decay into subclusters.

Seeing a research topic as a shared focus on scientific knowledge suggests that not separation but cohesion of views on knowledge should be the defining property of topics. We have tried to weaken this argument by pointing to core-periphery structures and by proposing the simple CPLC algorithm that constructs such structures inside well-separated link clusters (Havemann et al., 2019). This approach still rests on the assumption that topics can be represented by well separated clusters. Experiments with PsiMinL show that there are such topics but they do not prove that all research topics can be separated from the rest of a citation network. In dense cores of the network, separation could fail, as the occurrence of a *terra incognita* (a huge central cluster without substructures) in the analysis of astronomy and astrophysics seems to suggest (Havemann et al., 2017, p. 1105).

Technically, PsiMinL is an evolutionary algorithm that searches for local minima in the cost landscape with evaluation function Ψ. PsiMinL starts memetic evolutions from seed subgraphs, but the same valid cluster can be reached from different seeds (cf. Figure 2). In this sense, the cluster solution is independent of seeds. The construction of seeds influences only the time needed for a solution and its completeness.

All of the technical parameters of PsiMinL also do not affect the results but only the time needed to obtain them. The only numerical parameter that influences the shape of the clusters is the resolution r. In this study, I have tested a procedure that makes the results less dependent on r. I started with low r and then iteratively used the clusters as seed subgraphs for running PsiMinL for higher levels of r. Because lower r means a faster search, this strategy could also be advantageous when results at only one resolution level are needed.

Evolutionary algorithms on large networks need much computing time. PsiMinL, as with other algorithms, shifts the time problem at least partly to one of computing power by applying highly parallel procedures. Genetic operators can be applied parallelly on all individuals of a population. Because clusters are evaluated independently we can start PsiMinL parallelly from different seeds. Further optimization of PsiMinL could be reached by finding optimal sets of technical parameters such as population size and mutation rate. Another technique for reducing computing time could be to start with only 2 years and then use the resulting clusters as seeds for larger periods, similar to reducing a large graph by random sampling (Azaouzi, Rhouma, & Ben Romdhane, 2019, p. 23).

Finding a minimum in a large and rough cost landscape by applying an evolutionary strategy never comes to an end because we cannot prove that there is no lower place than the one found. PsiMinL searches for local minima and accepts a link cluster L as a valid solution if it is not made invalid by a lower place inside a radius of $r|L|$ in the landscape. That means that we cannot exclude that there are better variants of clusters, but we also cannot maintain that we have found all valid clusters. Sometimes, PsiMinL invalidates a cluster not in the first trials. That means that we cannot be sure that a found cluster is really valid, but we can at least assume a weak validity when PsiMinL is not able to find a path to a better cluster after several trials.

Applying PsiMinL for finding link clusters in citation networks needs preprocessing (data cleaning, construction of seeds)[16] and postprocessing (selection of valid solutions, finding cohesive cores). Running PsiMinL many times for many seeds requires not only computing time and power but also a clear organization of all procedures, selections, and validations. PsiMinL

[16] But note that tedious cleaning of citation data can be reduced to highly cited sources when citation networks of concept symbols are clustered.

cannot be recommended for a user only interested in results before the whole procedure is transformed into a routine of automatic actions. More experience is needed for optimizing the exploration of cost landscapes with PsiMinL. Then, we hope, we can make a step further in codifying the procedure.

ACKNOWLEDGEMENTS

As a member of the project team, Felix Mattes made all downloads and developed the algorithm for reference identification. Lixue Lin-Siedler helped in classifying references as scholarly ones. The team members and experts in IR, Thomas Risse, Wiebke Wemheuer-Vogelaar, and Mathis Lohaus commented on results and classified the 300 highly cited sources. Special thanks to Jochen Gläser who gave valuable advice on the whole process of data collection and processing. The memetic algorithm was implemented as an R-package by Andreas Prescher. I thank Michael Heinz for many discussions and for applying an alternative clustering method. He, Jochen Gläser, Alexander Struck, Mathis Lohaus, and Martin Enders also commented on drafts of the paper. The comments of two anonymous reviewers were also very helpful for improving the paper, many thanks! Finally I thank the developers of L^AT_EX and of R.[17]

COMPETING INTERESTS

The author has no competing interests.

FUNDING INFORMATION

This work is part of the Global Pathways project sponsored by DFG (grant RI 798/11-1).[18] Algorithm and R-package PsiMinL were developed in a project funded by the German Research Ministry (BMBF grant 01UZ0905).

DATA AVAILABILITY

The raw data used in this paper were obtained from the WoS database produced by Clarivate Analytics. Due to license restrictions, the data cannot be made openly available. To obtain WoS data, please contact Clarivate Analytics.[19]

The results of cleaning and clustering can be found on Zenodo (Havemann, 2020).

REFERENCES

Ahn, Y., Bagrow, J., & Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, *466*(7307), 761–764. **DOI:** https://doi.org/10.1038/nature09182, **PMID:** 20562860

Azaouzi, M., Rhouma, D., & Ben Romdhane, L. (2019). Community detection in large-scale social networks: State-of-the-art and future directions. *Social Network Analysis and Mining*, *9*(1), 23. **DOI:** https://doi.org/10.1007/s13278-019-0566-x

Butts, C. T. (2016). *sna: Tools for social network analysis*. Version 2.4. http://dk.archive.ubuntu.com/pub/pub/cran/web/packages/sna/sna.pdf

Chalupa, D., Hawick, K., & Walker, J. (2018). Hybrid bridge-based memetic algorithms for finding bottlenecks in complex networks. *Big Data Research*, *14*, 68–80. **DOI:** https://doi.org/10.1016/j.bdr.2018.04.001

Enders, M., Havemann, F., Ruland, F., Bernard-Verdier, M., Catford, J. A., … Jeschke, J. M. (2020). A conceptual map of invasion biology: Integrating hypotheses into a consensus network. *Global Ecology and Biogeography*, *29*(6), 978–991. **DOI:** https://doi.org/10.1111/geb.13082

Evans, T. S., & Lambiotte, R. (2009). Line graphs, link partitions, and overlapping communities. *Physical Review E*, *80*(1), 16105. **DOI:** https://doi.org/10.1103/PhysRevE.80.016105, **PMID:** 19658772

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*, 75–174. **DOI:** https://doi.org/10.1016/j.physrep.2009.11.002

[17] https://www.r-project.org.

[18] http://t1p.de/globalpathways.

[19] https://clarivate.com/products/web-of-science.

Gabardo, A. C., Berretta, R., & Moscato, P. (2020). M-link: A link clustering memetic algorithm for overlapping community detection. *Memetic Computing*, *12*, 87–99. **DOI:** https://doi.org/10.1007/s12293-020-00300-x

Garfield, E. (1985). History of citation indexes for chemistry: A brief review. *Journal of Chemical Information and Computer Sciences*, *25*(3), 170–174. **DOI:** https://doi.org/10.1021/ci00047a007

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. **DOI:** https://doi.org/10.1007/s11192-017-2296-z

Gläser, J., Heinz, M., & Havemann, F. (2015). Epistemic diversity as distribution of paper dissimilarities. In A. A. Salah, Y. Tonta, A. A. Akdag Salah, C. Sugimoto, & U. Al (Eds.), *Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference* (pp. 1006–1017). https://www.researchgate.net/publication/280298026_Epistemic_Diversity_as_Distribution_of_Paper_Dissimilarities

Havemann, F. (2020). Topics in research on international relations as clusters of citation links. **DOI:** https://doi.org/10.5281/zenodo.4181930

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. **DOI:** https://doi.org/10.1007/s11192-017-2302-5

Havemann, F., Gläser, J., & Heinz, M. (2019). Communities as well separated subgraphs with cohesive cores: Identification of core-periphery structures in link communities. In L. M. Aiello, C. Cherifi, H. Cherifi, R. Lambiotte, P. Lió, & L. M. Rocha (Eds.), *Complex networks and their applications VII*, Studies in Computational Intelligence (pp. 219–230). Cham: Springer. **DOI:** https://doi.org/10.1007/978-3-030-05411-3_18

Havemann, F., Gläser, J., Heinz, M., & Struck, A. (2012). Identifying overlapping and hierarchical thematic structures in networks of scholarly papers: A comparison of three approaches. *PLOS ONE*, *7*(3), e33255. **DOI:** https://doi.org/10.1371/journal.pone.0033255, **PMID:** 22479376, **PMCID:** PMC3314014

Kristensen, P. M. (2018). International relations at the end: A sociological autopsy. *International Studies Quarterly*, *62*(2), 245–259. **DOI:** https://doi.org/10.1093/isq/sqy002

Lu, Z., Hao, J.-K., & Wu, Q. (2020). A hybrid evolutionary algorithm for finding low conductance of large graphs. *Future Generation Computer Systems*, *106*, 105–120. **DOI:** https://doi.org/10.1016/j.future.2019.12.049

Marshakova, I. V. (1973). Sistema svyazey mezhdu dokumentami, postroyennaya na osnove ssylok (po ukazatelyu "Science Citation Index"). *Nauchno-Tekhnicheskaya Informatsiya Seriya 2 – Informatsionnye Protsessy i Sistemy*, *6*, 3–8.

Neri, F., Cotta, C., & Moscato, P. (Eds.) (2012). *Handbook of memetic algorithms*. Berlin: Springer. **DOI:** https://doi.org/10.1007/978-3-642-23247-3

Pizzuti, C. (2017). Evolutionary computation for community detection in networks: A review. *IEEE Transactions on Evolutionary Computation*, *22*(3), 464–483. **DOI:** https://doi.org/10.1109/TEVC.2017.2737600

Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., & Parisi, D. (2004). Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, *101*, 2658–2663. **DOI:** https://doi.org/10.1073/pnas.0400054101, **PMID:** 14981240, **PMCID:** PMC365677

Risse, T., Wemheuer-Vogelaar, W., & Havemann, F. (2020). Theory makes global IR hang together. Lessons from citation analysis. Preprint: http://dx.doi.org/10.17169/refubium-28510

Rosvall, M., Delvenne, J.-C., Schaub, M. T., & Lambiotte, R. (2019). Different approaches to community detection. In P. Doreian, V. Batagelj, & A. Ferligoj (Eds.), *Advances in network clustering and blockmodeling* (pp. 105–119). Chichester: John Wiley & Sons. **DOI:** https://doi.org/10.1002/9781119483298.ch4

Small, H. G. (1978). Cited documents as concept symbols. *Social Studies of Science*, *8*(3), 327–340. **DOI:** https://doi.org/10.1177/030631277800800305

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*, 265–269. **DOI:** https://doi.org/10.1002/asi.4630240406

Small, H., & Sweeney, E. (1985). Clustering the Science Citation index® using co-citations. *Scientometrics*, *7*(3), 391–409. **DOI:** https://doi.org/10.1007/bf02017157

Tang, W., Zhao, L., Liu, W., Liu, Y., & Yan, B. (2019). Recent advance on detecting core-periphery structure: A survey. *CCF Transactions on Pervasive Computing and Interaction*, *1*(3), 175–189. **DOI:** https://doi.org/10.1007/s42486-019-00016-z

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. **DOI:** https://doi.org/10.1162/qss_a_00035