A Link-based Memetic Algorithm for Reconstructing Overlapping Topics from Networks of Papers and their Cited Sources

Frank Havemann[1], Jochen Gläser[2] and Michael Heinz[1]

[1] *frank.havemann@ibi.hu-berlin.de* [1] *michael.heinz@rz.hu-berlin.de*
Humboldt-Universität zu Berlin, Berlin School of Library and Information Science, Dorotheenstraße 26, 10099 Berlin (Germany)

[2] *jochen.glaser@ztg.tu-berlin.de*
TU Berlin, Center for Technology and Society, Hardenbergstr. 16-18, D-10623 Berlin (Germany)

Abstract

In spite of recent advances in field delineation methods, enduring problems such as the impossibility to justify necessary thresholds and the difficulties in comparing thematic structures obtained by different algorithms leave bibliometricians with a sense of uneasiness about their methods. In this paper, we propose and demonstrate a new approach to the delineation of thematic structures that attempts to fit the methods for topic delineation to the properties of topics. We derive principles of topic delineation from a theoretical discussion of thematic structures in science. Applying these principles, we cluster citation links rather than publication nodes, use predominantly local information and grow communities of links from seeds in order to allow for pervasive overlaps of topics. The complexity of the clustering task requires the application of a memetic algorithm that combines probabilistic evolutionary strategies with deterministic local searches. We demonstrate our approach by applying it to a network of 14,954 Astronomy & Astrophysics papers and their cited sources.

Conference Topic

Methods and techniques (special session on algorithms for topic detection)

Introduction

The identification of thematic structures (topics or fields) in sets of papers is one of the recurrent problems of bibliometrics. It was deemed one of the challenges of bibliometrics by van Raan (1996) and is still considered as such despite the significant progress and a plethora of methods available. Major developments since van Raan's paper include approaches that cluster the whole Web of Science based on journal-to-journal citations, co-citations, or direct citations, the advance of hybrid approaches that combine citation-based and term-based techniques, and term-based probabilistic methods (topic modelling). However, methodological problems endure and leave bibliometricians with a sense of uneasiness about their methods. Advanced methods still apply thresholds that must be arbitrarily set and adapted to the specific structures that shall be obtained. The relevance of the structures identified by bibliometric methods are difficult to verify independently, and the relationships between thematic structures are difficult to assess. A recent analysis by Hric et al. (2014) found that current algorithms for the detection of communities in network of papers respond to topological properties of networks but not necessarily to the underlying real-world properties of nodes clustered. This observation casts further doubts on the fundamental assumption underlying bibliometric methods for topic delineation, namely that the topics reconstructed using *structural properties of networks of papers* reflect *thematic properties of the research published in those papers*.

In this paper, we propose and demonstrate a new approach to the delineation of thematic structures. We derive principles of topic delineation and criteria for the assessment of algorithms from a theoretical discussion of properties of thematic structures in science. Applying these principles, we cluster citation links rather than publication nodes, use predominantly local information, and grow communities from seeds in order to allow for pervasive overlaps of topics. The complexity of the clustering task requires the application of a memetic algorithm that combines nondeterministic evolutionary strategies with deterministic local searches. We demonstrate our approach by applying it to a network of 14,954 Astronomy & Astrophysics papers and their cited sources.

Strategy, Methods and Data

Theoretical considerations and strategy

We define topics as theoretical or empirical knowledge about objects or methods of research that is a common focus for a set of research processes because it provides a reference for the decisions of researchers – the formulation of problems, the selection of methods or objects, the organisation of empirical data, or the interpretation of data (on the social ordering of research by knowledge see Gläser 2006). This definition resonates with Whitley's (1974) description of research areas but abandons the assumption that topics form a hierarchy. It only demands that some scientific knowledge is perceived similarly by researchers and influences their decisions.

This weak definition is linked to three properties of topics that create the problems for bibliometrics:

1) The fractal nature of knowledge has been described by van Raan (1991) and Katz (1999). Topics can have any 'size' (however measured) between the smallest (emerging topics that just concern one researcher) and very large thematic structures (fields or even themes cutting across several fields). Methods for topic identification should thus not be biased against any particular topic size.

2) Given the multiple objects of knowledge that can serve as common reference for researchers, topics inevitably overlap. Publications commonly contain several knowledge claims, which are likely to address different topics (Cozzens, 1985; Amsterdamska & Leydesdorff, 1989). Methods for topic identification should thus take into account that bibliometric objects (publications, authors, journals, and cited sources) are likely to belong to several topics simultaneously. Methods also should enable the reconstruction of topics that overlap pervasively (i.e. not only in their boundaries).

3) All topics emerge from coinciding autonomous interpretations and uses of knowledge by researchers (see e.g. the case studies discussed by Edge and Mulkay, 1976, pp. 350-402). While individual researchers may launch topics and advocate them, the latter's content and fate depends on the ways in which they are used by others. From this follows that topics are local in the sense that they are primarily topics to the researchers whose decisions are influenced by and who contribute to them. Methods for topic identification can reconstruct this insider perspective by using local information. Global approaches create different representations of topics by finding a compromise between insider perspectives and all outsider perspectives on topics.

Methods

For a detailed description of the method see Havemann, Gläser, & Heinz (2015). We operationalise 'topic' as a set of thematically related papers but cluster citation links instead of papers because the former can be assumed the thematically most homogenous bibliometric objects (see Evans & Lambiotte, 2009; and Ahn, Bagrow & Lehmann, 2010 on link clustering).

Cost Function: We followed the suggestion by Evans and Lambiotte (2009) to obtain link clusters by clustering vertices in a network's line graph and defined a local cost function $\Psi^*(L)$ of link set L in the line-graph approach. The internal degree k_i^{in} (L) of node i is defined as the number of links in L attached to i. The external degree of a node is obtained by subtracting the internal from the total degree: $k_i^{\mathrm{out}}(L) = k_i - k_i^{\mathrm{in}}(L)$. External degrees k_i^{out} are weighted with subgraph membership-grade k_i^{in}/k_i of boundary node i to obtain a measure of external connectivity of link set L:

$$\sigma(L) = \sum_{i=1}^{n} \frac{k_i^{out}(L)k_i^{in}(L)}{k_i} \quad (1)$$

where n is the number of all nodes. The sum can be restricted to boundary nodes because only for boundary nodes of L is $k_i^{\mathrm{out}}k_i^{\mathrm{in}} > 0$. A simple size normalization that accounts for the finite size of the network is achieved by adapting the ratio cut suggested by Wei and Cheng (1989) for link communities, which leads us to the cost function *ratio node-cut* $\Psi^*(L)$:

$$\Psi^*(L) = \frac{\sigma(L)}{k_{in}(L)(1 - \frac{k_{in}(L)}{2m})} \quad (2)$$

where m is the number of all links and $k_{in}(L)$ is the sum of all internal degrees $k_i^{\mathrm{in}}(L)$. $\Psi^*(L)$ essentially relates external to total connectivity of link set L. It can be used to identify link communities (sets of links that are well connected internally and well separated from the rest of the graph) by finding local minima in the cost landscape.

Since the cost landscape is often very rough—has many local minima that sometimes correspond to very similar subgraphs—the resolution of the algorithm must be defined by setting a minimum distance (number of links that differ) between subgraphs corresponding to different local minima. We define the range of a community as the environment in which no subgraph exists that has a lower Ψ^* value. For our experiments with the citation network of astrophysical papers we set a community's minimum range at one third of its size.

Algorithm: The cost function Ψ^* is used in a clustering algorithm that grows communities from seeds. This approach fulfils two more principles derived from our definition of a topic. The independent construction of each community prevents a size bias of the algorithm and enables pervasive overlaps.

```
choose a connected subgraph as a seed
initialize population P by mutating the seed with high variance several times and adapt mutants
while the best community is not too old do
    mutate the best community with low variance and adapt the mutants
    if a mutant is new and its cost is lower than highest cost then
        add it to population P
    end if
    cross the best community with other communities and adapt the offspring
    if offspring is new and its cost is lower than highest cost then
        add it to population P
    end if
    select the best individuals so that the population size remains constant
    if there is no better best community for some generations and innovation rate is low then
        renew the population (mutate the best community with high variance and adapt it)
        select the best individuals so that the population size remains constant
    end if
end while
```

Figure 1. Pseudocode of memetic evolution.

The task of finding communities in large networks is always very complex and requires the use of heuristics. We chose a memetic algorithm that accelerates the search by combining non-deterministic evolution with a deterministic local search in the cost landscape (Neri, Cotta, & Moscato, 2012). In our algorithm, populations of subgraphs evolve because after a random initialization of a population of some definite size, the genetic operators of crossover, mutation, and selection are repeatedly applied (Fig. 1). Each crossover and mutation is followed by a local search.

Data

The algorithm is applied to the citation network of 14,954 papers published 2010 in 53 journals listed in the category Astronomy & Astrophysics of the Journal Citation Reports 2010 (the journal *Space Weather* with 45 articles was accidentally left out). We downloaded all articles, letters and proceedings papers from the Web of Science. Reference data had to be standardised with rule-based scripts. To reduce the complexity of the network, we omitted all sources that are cited only once because they do not link papers and their removal should not unduly influence clustering. We excluded 184 papers that are not linked to the giant component of the citation network and proceeded with a network of 119,954 nodes that are connected by 536,020 citation links. We neglected the direction of citation links and analysed an undirected unweighted connected graph.

Experiments and Preliminary Results

Constructing the seed population

Since topics can assume all possible sizes, the algorithm should start from differently sized seed graphs. In our experiments, we combined two strategies for obtaining seeds. First, we used Ward clustering with a similarity measure derived from theoretical considerations (Gläser, Heinz & Havemann, 2015). We ordered all hard clusters by their stability (the length of their branch in the dendrogram) and selected the most stable but not too large clusters (a total of 63) as seeds. In addition, we used the citation links of 969 randomly selected papers as seed graphs.

Each seed was first adapted by a local search and then used to initialise the population of 16 different communities by mutating the seed with a variance of 15%.

Owing to the randomness of the evolutionary mechanisms the choice of seed graphs is unlikely to affect the clustering results. However, it is likely to effect the efficiency of the algorithm.

Running the memetic algorithm

Up to ten experiments were run with each seed. The standard mutation variance in each experiment was 5%, i.e. up to 5% of the nodes were randomly exchanged. The variance was increased to 15% for one mutation if Ψ^* values did not improve for 10 generations. Again, we assume these parameters to effect the algorithm's efficiency rather than its outcomes.

Table 1. Examples of experiments with the memetic algorithm.

Community	*Seed sub-graph*		*Number of generations*	*Community*		*Remaining nodes from seed*
	Size	*Ψ^* value*		*Size*	*Ψ^* value*	
1	13,469	.0692	339	10,586	.0339	10,380
2	19,697	.1174	233	35,159	.0397	18,860
3	35	.4075	232	33	.0047	0
4	76	.5498	203	28	.0975	0

Experiments with the seeds described above resulted in a total of 3,944 distinct communities, 1,375 of which were disregarded because there were better communities within a distance of less than one third of their size. The remaining 2,569 communities were ordered by increasing *Ψ** values. Table 1 provides exemplary descriptions of some of the experiments. We then calculated the relative coverage of the network as a function of *Ψ** by successively uniting the *L*-sets of the ranked communities. Relative coverage is the ratio of the union's size to the number of all links *m* (Fig. 2). This function has a sharp bend at *Ψ**=0.10458, shortly below maximum coverage. We used this *Ψ** value as cutoff point, which gives us a preliminary result of 154 communities that cover 98.9 % of all links.

Currently, each of these 154 best communities is used as a seed for a refined local search that adds or removes single links instead of nodes with all their links. For some of the 154 communities this additional local search has already led to better communities.

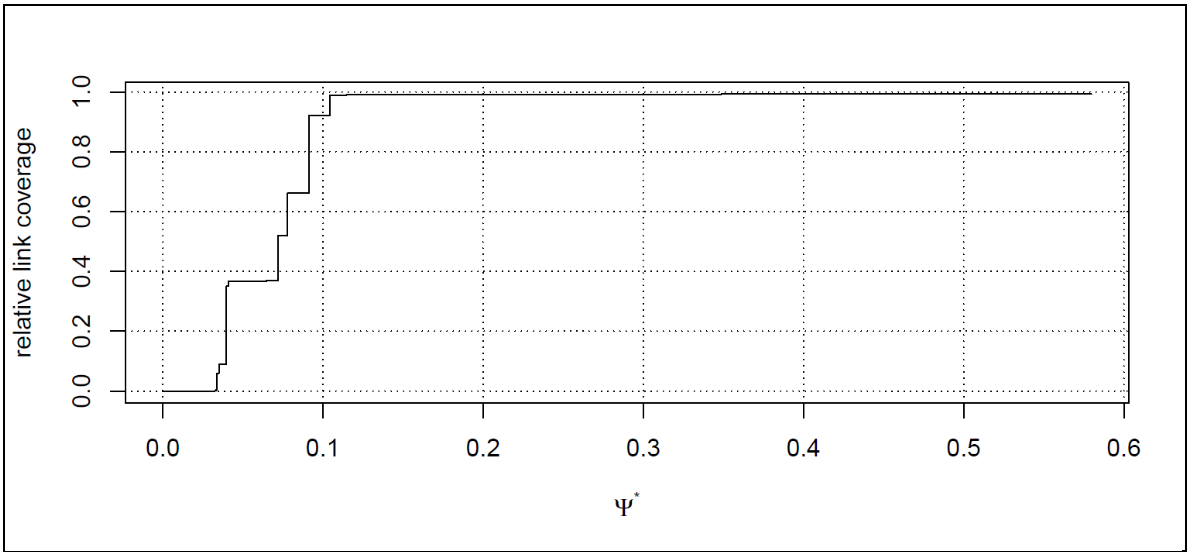

Figure 2. Relative coverage of the network by communities as a function of a Ψ* threshold.

Preliminary results

The 154 communities vary in their size between 9 and 49,324 nodes. Some of the communities overlap pervasively. Seventy communities were not a subset of any other community. The other 84 communities were subsets of one (12 communities) to 28 other communities (1). In Figure 3 we plot sizes and cost of the 154 best communities. Blue circles represent communities that are subsets of others. Green circles represent communities that overlap with another community in 95% of their nodes. All other communities are represented by red circles. The numbers in four circles refer to the communities described in Table 1.

The communities form a poly-hierarchy because some smaller communities are subsets of two larger communities that have no hierarchical subset relation. A community can also have a rest of nodes which are not members of any of its sub-communities.

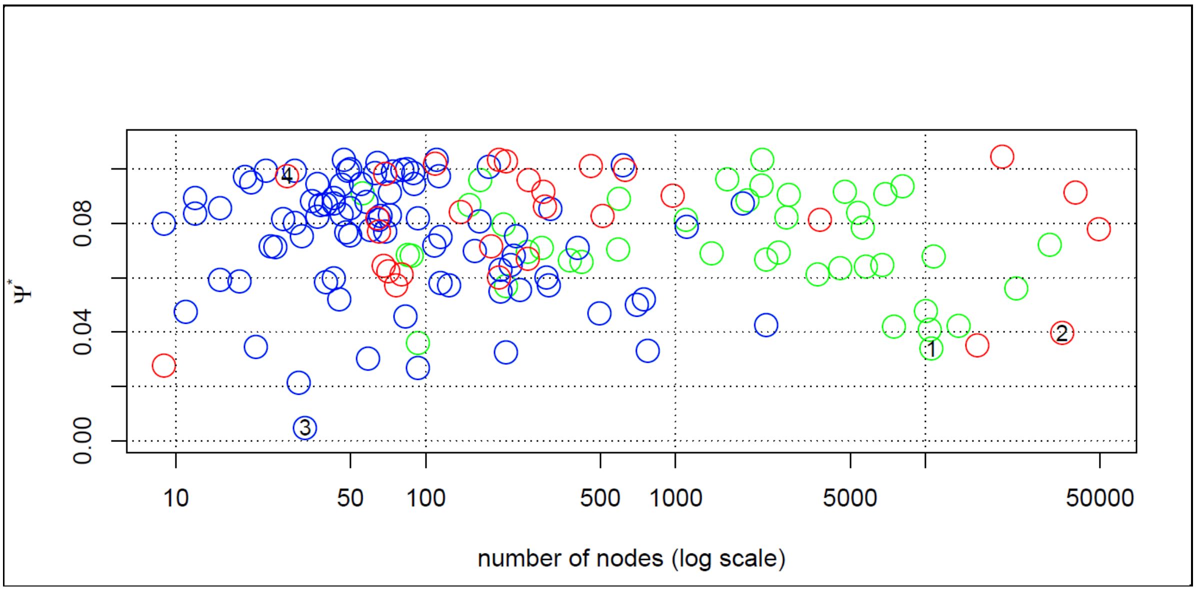

Figure 3. Sizes and Ψ* values of a set of communities covering 98.9% of the graph.

Conclusions

The communities have the structural properties of topics that were derived from the definition. Comparisons with other cluster solutions and tagging of communities will show whether the communities are consistent. We will test the dependence of results on parameter and seed choice with a smaller network. Ultimately, only a discussion with experts can show whether the communities obtained provide one of the possible scientifically meaningful cluster solutions of the astronomy and astrophysics dataset.

Acknowledgments

The work published here was funded by the German Research Ministry (01UZ0905). We thank Andreas Prescher for programming a fast C++-based R-package for parallel node-wise memetic search in the Ψ^* - landscape.

References

Ahn, Y., Bagrow, J. & Lehmann, S. (2010). Link communities reveal multiscale complexity in networks. *Nature*, *466*(7307), 761-764.

Amsterdamska, O. & Leydesdorff, L. (1989). Citations: Indicators of Significance? *Scientometrics*, 15, 449-471.

Cozzens, S. E. (1985). Comparing the Sciences: Citation Context Analysis of Papers from Neuropharmacology and the Sociology of Science. *Social Studies of Science, 15*(1), 127-153.

Edge, D. & Mulkay, M. J. (1976). *Astronomy Transformed: The Emergence of Radio Astronomy in Britain*. New York: John Wiley & Sons, Inc.

Evans, T. & Lambiotte, R. (2009). Line graphs, link partitions, and overlapping communities. *Physical Review E*, *80*(1), 16105.

Gläser, J. (2006). *Wissenschaftliche Produktionsgemeinschaften. Die soziale Ordnung der Forschung.* Frankfurt a. M.: Campus.

Gläser, J., Heinz, M. & Havemann, F. (2015). Measuring the diversity of research. Paper submitted to the 15th International Conference on Scientometrics and Informetrics, Istanbul, 29 June -4 July 2015.

Havemann, F., Gläser, J. & Heinz, M. (2015). Detecting Overlapping Link Communities by Finding Local Minima of a Cost Function with a Memetic Algorithm. Part 1: Problem and Method. arXiv:1501.05139.

Hric, D., Darst, R. K. & Fortunato, S. (2014). Community detection in networks: Structural communities versus ground truth. *Physical Review E*, *90*, 062805.

Katz, J. S. (1999). The self-similar science system. *Research Policy*, *28*, 501-517.

Neri, F., Cotta, C. & Moscato, P. (Eds.) (2012). *Handbook of Memetic Algorithms*, Volume 379 of Studies in Computational Intelligence. Berlin: Springer.

Van Raan, A. F. J. (1991). Fractal Geometry of Information Space as Represented by Co-Citation-Clustering. *Scientometrics*, *20*, 439-449.

Van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, *36*, 397-420.

Wei, Y.-C. & Cheng, C.-K. (1989). Towards efficient hierarchical designs by ratio cut partitioning. In *IEEE International Conference on Computer-Aided Design, 1989. ICCAD-89. Digest of Technical Papers*, pp. 298–301.

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