A local cohesion-maximising algorithm for the exploration of publication networks

Matthias Held[*], Bastian Steudel,** Jochen Gläser[*]

[*]*matthias.held@tu-berlin.de; jochen.glaser@tu-berlin.de*
Social Studies of Science and Technology, TU Berlin, Germany
***Bastian.Steudel@gmx.net*

The dominant approach to the reconstruction of scientific topics in networks of publications is based on the application of global community detection algorithms. However, some properties of these algorithms are at odds with the sociological understanding of topics. We present for consideration a new local bibliometric algorithm which is in line with sociological definitions of topics and reconstructs dense regions in bibliometric networks locally.

1. Introduction

Reconstructing scientific topics from networks of papers is a primary focus of bibliometrics, with applications in both science studies and science policy. The dominant approach applies global algorithms – algorithms that partition the whole network by optimising a quality function to obtain a partition – with data models based on direct citation or bibliographic coupling and interprets the resulting clusters as topics (Gläser et al., 2017). The most popular global algorithms prioritise the separation of clusters over their coherence (Held, 2022).

This approach is problematic because it inadvertently decouples the bibliometric reconstruction of topics from the sociological discussion of their role in the production of scientific knowledge. Sociological definitions of topics imply an emphasis on coherence and on the perspective of those contributing to the topic, i.e. on a local perspective. Not surprisingly, established bibliometric approaches to topic reconstruction proved unsuccessful when applied to 'ground truths' that were defined by scientists (Held et al., 2021).

These problems and their apparent roots in currently popular algorithms in bibliometrics motivated our search for new algorithms that might be more conducive to a bibliometric operationalisation of the sociological concept of 'topic'. Local algorithms – algorithms that grow clusters from seed subgraphs until a condition for their termination is met – appear to be a promising solution to the tenets of using local information and of allowing for overlapping topics. Some of these algorithms use quality functions that maximise cohesion, which corresponds to the sociological understanding of topics as shared perspectives of researchers.

In this paper, we provide a rationale for using local cohesion-maximising algorithms, present such an algorithm and discuss one application.

2. A rationale for local density-maximising algorithms

2.1 Theory

An important methodological starting point of our search for approaches to topic reconstruction is the demand that these approaches, as procedures of empirical identification, operationalise a theoretical concept. We follow Havemann et al. (2017: 1091) in defining a topic 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".

The researchers who share such a frame of reference form a scientific specialty or scientific community, i.e. a collective that jointly advances the shared knowledge and has a collective identity (self-perception) of jointly advancing that knowledge (Gläser, 2019; Whitley, 2000). This joint activity is based on intense communication because community members'

publications contain contributions that are offered to fellow community members for further use (Kuhn, 1970 [1962]: 19, 23, 177). It is also based on, and strengthens, the thematic similarity of community members' work. Finally, from the definition of a topic as a joint frame of reference follows that a topic is first and foremost a topic to those who work on it. The insider perspective that constitutes a topic is likely to deviate from outsider perspectives, i.e. perspectives of colleagues from other communities.

These theoretically derived properties of topics should correspond to properties of subgraphs in publication networks if these subgraphs are meant to represent topics. Dense communication should be reflected in above-average subgraph cohesion in direct citation networks. Thematic similarity should be reflected in above-average subgraph density in bibliographic coupling networks. The insider perspective should be realised by the use of local information (information about a subgraph and its environment) for its delineation (Held, 2022). Taken together, these operationalisations suggest experimenting with local density-maximising algorithms, which can be applied to traditional bibliometric data models like direct citation and bibliographic coupling.

2.2 Local algorithms

In network research, many different *local community detection algorithms* (LCDA), i.e. algorithms which start locally from a seed in a network and grow a so-called community around it, have been developed (Dilmaghani et al., 2021). LCDAs share the classical idea of a community in a network having "more edges 'inside' [...] than edges linking vertices […] with the rest of the graph" (Fortunato, 2010).

This understanding of communities in networks calls for a maximisation of a subgraph's cohesion and separation. While global network partition algorithms must solve this problem by striking a compromise because in a partition neither can be optimised individually (Fortunato & Hric, 2016), local algorithms can focus on either separation or cohesion. Most local algorithms evaluate a community's quality by its separation from its environment, which is frequently measured as conductance (outward edges divided by volume, Hamann et al., 2017) or local modularity (Clauset, 2005: 2). Only few local algorithms maximise cohesion. These include, among others,

- the Local Tightness Expansion algorithm (LTE), which grows the subgraph by adding nodes that increase the subgraph's "tightness" (shared neighbours of nodes inside the subgraph compared to neighbours of nodes inside and outside of the subgraph) and also uses "tightness" as termination criterion (Huang et al., 2011), and
- the Triangle Based Community Expansion (TCE) algorithm (Hamann et al., 2017), which adds nodes when they have a large share of triangular relationships with the subgraph compared to the nodes' degree but uses a separation-oriented criterion (conductance) for termination.

While some algorithms find their own seed to start from (Dilmaghani et al., 2021b: 762), e.g. by random selection, others have to be provided a user-defined seed. Some algorithms use user-defined seeds as starting point for finding a suitable seed in its surrounding, e.g. by searching for clique(-like) structures that include the seeds (Fanrong et al., 2014) or degree-central nodes (Q. Chen et al., 2013). Others start the expansion directly from the user-defined seed.

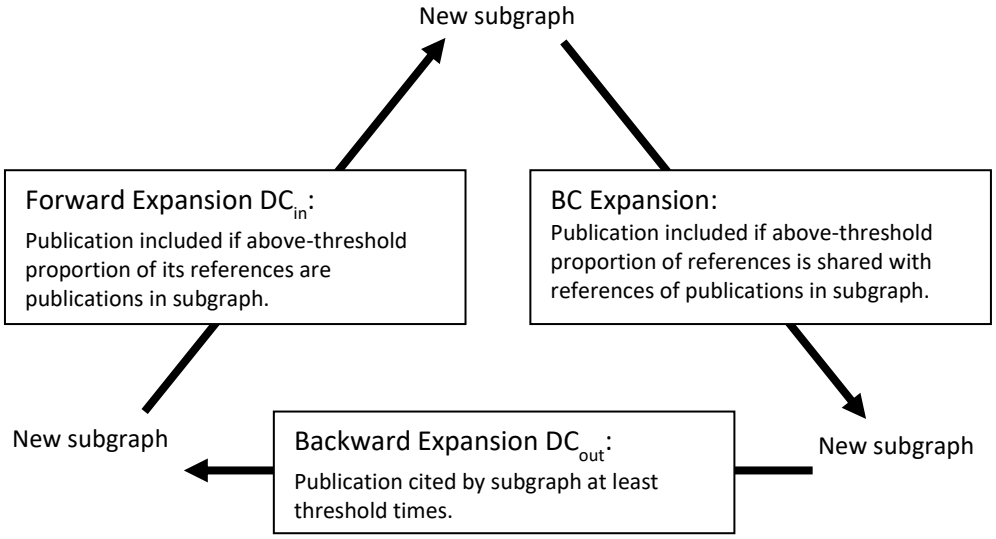
To our knowledge, only two attempts have been made to utilise local algorithms for *bibliometric* questions. Havemann et al. (2017) used a separation-based memetic local algorithm for topic reconstruction. C. Chen (2018) proposed cascading citation expansion, which is a local approach but not an algorithm for community detection.

2.3 The MALBA algorithm

We present for consideration a Multilayer Adjustable Local Bibliometric Algorithm (MALBA), which is inspired by the LTE algorithm. MALBA constructs cohesive communities in networks

of papers by iteratively growing a subgraph from a seed, i.e. it operates *locally*. Publications are added to the subgraph if they are densely connected in at least one of the two data models direct citation or bibliographic coupling, i.e. it operates in a *multi-layered* network (Figure 1). It can also be applied to networks based on only one of the two data models. The thresholds for the density of connections are *adjustable* by the user. MALBA terminates when no more publications exist whose connections to the subgraph are above one of the density thresholds. The separation of the subgraph from its neighbourhood is considered only collaterally because papers that are not connected well enough to be included are in turn better separated from the subgraph.

Figure 1: Iterative expansion steps of MALBA.[1]



MALBA can be applied to pre-existing networks, from which subgraphs are selected as seeds. In this mode, MALBA can support the exploration of networks by identifying dense regions. Alternatively, the algorithm can be used to explore a publication database directly by starting from a seed and searching the database for densely connected publications. In this case, MALBA utilises all information about a subgraph's environment that exists in the database but provides less information about less well-connected publications. The interface works with both approaches. Figure 2 presents the pseudocode of MALBA.

[1] A change in the order of the three steps (e.g. BC → DCin → DCout, or DCin → BC → DCout) does not result in different subgraphs.

Figure 2: Pseudocode of MALBA

```
Input: Seed subgraph
Initialize: current_community C is the seed subgraph
while switch==True do
   DCout_step: current_community's cited pubications above DCout_threshold
    if new publications are found in DCout_step then
        add them to current_community C
        DCout_switch = True
    else
        DCout_switch = False

    BC_step: current_community's bibliographically coupled publications above BC_threshold
    if new publications are found in BC_step then
        add them to current_community C
        BC_switch = True
    else
        BC_switch = False

    DCin_step: publications citing current_community above DCin_threshold
    if new publications are found in DCin_step then
        add them to current_community C
        DCin_switch = True
    else
        DCin_switch = False

    if BC_switch == False and DCin_switch==False and DCout_switch==False then
        switch = False
```

The user can affect the operation of MALBA in three ways:
1) By deciding to work with a pre-existing network or to explore a database. We already mentioned the main differences between the two approaches.
2) By constructing a seed subgraph as starting point for MALBA. The seed has a strong influence on the subgraph both through its size and through the region of the network in which it is located.
3) By deciding on the thresholds. The interface offers the option of automatically identifying the thresholds that return the largest subgraph that can be grown out of the seed with the algorithm terminating (see 3. Experiments). However, the user can also set thresholds manually to achieve an earlier termination of the algorithm. Higher thresholds focus on reconstructing denser regions, lower thresholds also allow to reconstruct less dense regions.

Previous experiments with MALBA revealed the following common behaviours in bibliometric networks:
(1) There are combinations of thresholds for each of the connections at which the algorithm terminates with a subgraph that is much smaller than the network or the database because no new nodes can be added. Lower thresholds lead to an exponential growth of the subgraph until it covers the whole network or database.
(2) A minimum size of the seed (which depends on the region of the network) is necessary for the subgraph to grow at all.
(3) A subgraph grown from a seed can itself be used as a seed for further growth.

3. Experiments

We report first experiments with MALBA in which we explore publications on a topic from library and information science – the h-index. We chose this topic because it makes it easier to understand publications included in the subgraph and in its environment. We discuss the reasons

why publications may be included or excluded, the impact of seed sizes, and the impact of thresholds.
In the experiments presented in this paper, we used MALBA to explore directly the stable version from July 2022 of the bibliometric database provided by the German "Competence Network for Bibliometrics", which consists of Web of Science data. When processing publications indexed in this database, we excluded all non-source items because their influence on bibliographic coupling (thematic similarity) and citation (communication) cannot be unambiguously assessed. The ratio-based thresholds for DC_{in} and BC used by MALBA make it more difficult for publications with many non-source items in their reference lists to be included in the subgraph.

3.1 Seeds
The algorithm is started with a seed set of seven most highly cited bibliometric publications in the WoS that have "h-index" in their title (Figure 3).

Figure 3: Seven highly cited publications for the topic h-index used as seed.

#	Seed Publications	Citations
1	Jin et al. (2007). The R-and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, *52*(6), 855-863.	438
2	Bornmann et al. (2008). Are there better indices for evaluation purposes than the h index? […] *JASIST*, *59*(5), 830-837.	338
3	Alonso et al. (2009). h-Index: A review focused in its variants, computation […] fields. *Journal of Informetrics*, *3*(4), 273-289.	538
4	Bornmann/Daniel (2007). What do we know about the h index?. *JASIST*, *58*(9), 1381-1385.	345
5	Hirsch (2007). Does the h index have predictive power?. *PNAS*, *104*(49), 19193-19198.	640
6	Costas/Bordons (2007). The h-index: Advantages, limitations […] . *Journal of Informetrics*, *1*(3), 193-203.	309
7	Bar-Ilan (2008). Which h-index?—A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, *74*(2), 257-271.	446

If only the seminal paper by Hirsch from 2005 from which the h-index topic emerged (which is not shown in Figure 3) is used as seed, the subgraph does not grow at all. This is not surprising because the original Hirsch paper is a publication outside the field of bibliometrics. When only a subset of the seven publications in Figure 3 is used, the algorithm adds only few publications and terminates at a maximum of 12 publications (Fig. 4, left column).

3.2 Results
When starting from the seed consisting of the 7 publications in Figure 3, the subgraph terminates at 805 publications (Fig. 4, middle column), at the thresholds DC_{in}=0.55, BC=0.95, DC_{out}=11. This means that at each stage of growth, publications were added to the subgraph if at least 55% of their references (source items) were publications included in the subgraph, if they shared at least 95% of their references (source items) with references (source items) of the subgraph's publications, or if they were cited at least 11 times by the subgraph. Lowering any of these thresholds slightly (e.g. DC_{in} to 0.50) leads to an exponential growth of the subgraph without termination. Smaller subgraphs can be obtained by increasing the thresholds. When the seed size is further increased to 15-20 publications and the same thresholds are used, almost the same subgraph emerges.
In the surrounding of this subgraph, we find false negatives – publications that address the h-index but are not included (FNs, Fig. 4) – and true negatives (TNs). An example of an FN is the study by Montazerian et al. (2019) "A new parameter for (normalized) evaluation of H-index: countries as a case study". It has only 54% of its references in the subgraph and thus did not pass the DC_{in} threshold of 55% (nor DC_{out} or BC, as it is cited by the subgraph only 2 times and shares only 80% of its reference, respectively). The FNs demonstrate that any threshold is bound to create "near misses", i.e. that a definitive delineation of a topic is not possible.

Most of the publications that were not included in the subgraph were TNs, e.g. the paper by Kosmulski (2018) "Are you in top 1%(1‰)?", which has 45% of its references in the subgraph but is not a clear h-index publication. Another TN is the study by Abramo et al. (2013) "The importance of accounting for the number of co-authors and their order when assessing research performance at the individual level in the life sciences" which has only 6 of 21 references in the subgraph (and is not cited by the subgraph while sharing 81% of its references) and deals with the h-index only marginally.

We used the subgraph we obtained with the thresholds given above as seed for a second run of MALBA with thresholds DC_{in}=0.80, BC=0.90, DC_{out}=11. This led to a termination at 1,320 publications. After this increase by more than 500 publications, some FNs which were found after the first run are still FNs, for example Bertoli-Barsotti and Lando (2015) "On a formula for the h-index". The large increase in publications after the second run led to the inclusion of some previous FNs (the study by Montazerian et al., for example), but, however, also leads to the addition of false positives. For example, the abovementioned study by Abramo et al. (2013) is now included in the subgraph of 1,320 publications.

Figure 4: Varying the seed sizes leads to different terminal subgraphs.

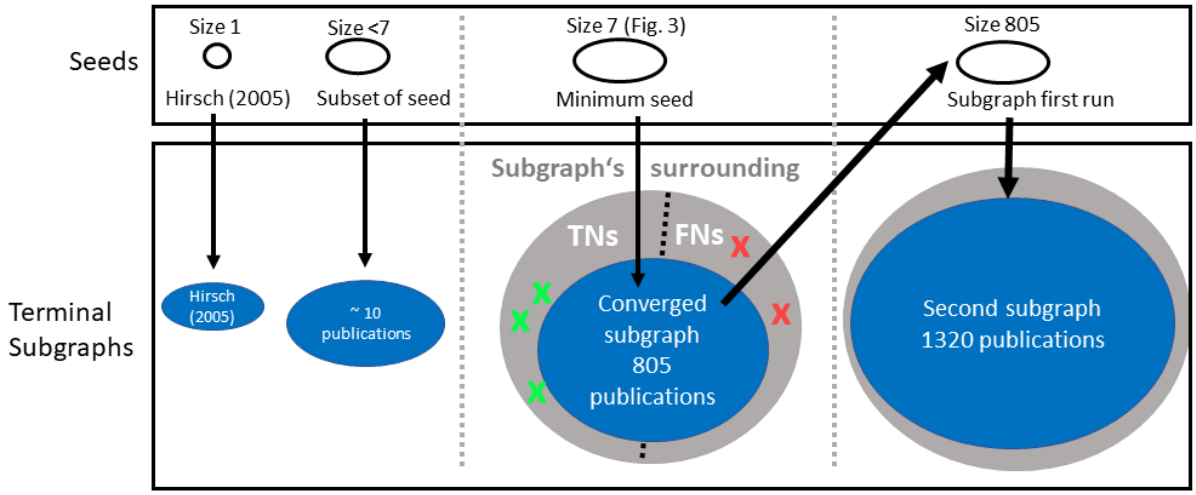
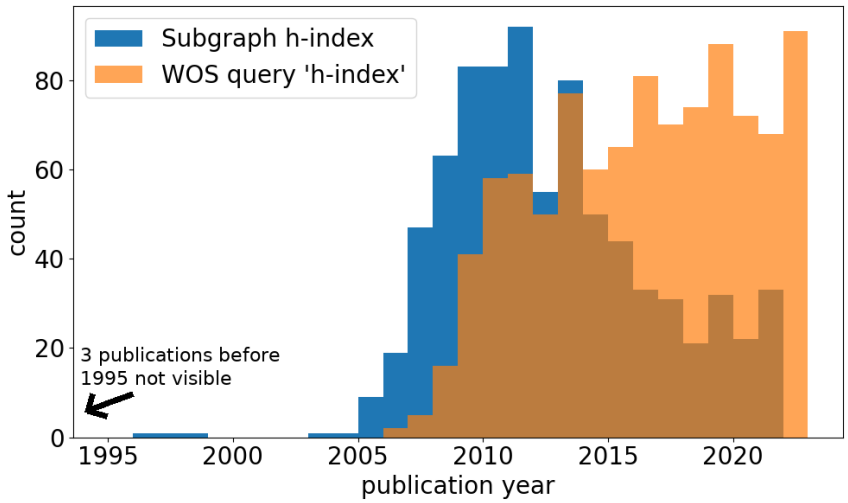


Figure 5 shows the distribution of the 805 publications over publication years and the number of publications with the keyword "h-index" for each year. The patterns clearly differ, which means that the growth of the subgraph is not influenced by the increasing number of publications on a topic.

Figure 5: Number of publications per year in MALBA subgraph and lexical query 'h-index'.



4. Discussion and future work

Local algorithms like MALBA are fully transparent because for every publication, the reason why it is included in a subgraph can be identified. This makes it possible to explore the match of subgraphs and their environment thematically and to identify true/false positives and true/false negatives with regard to the reconstruction of a topic.

While the reconstruction of the h-index topic with MALBA looks promising, we cannot claim yet that MALBA is suitable for reconstructing topics. Further experiments are necessary, including:

- a further exploration of subgraphs of h-index publications and their environments;
- experiments with MALBA and only one data model (direct citation or bibliographic coupling);
- a validation of MALBA with ground truths like the ones used in Held et al. (2021); and
- comparisons between the exploration of pre-existing networks and the exploration of publication databases.

Competing interests

The authors have no competing interests.

Funding information

This work was partly supported by the German Ministry of Education and Research (Grant 16PU17003).

Open science practices

We plan to make MALBA fully accessible to other researchers at the time of the conference. MALBA currently interrogates a proprietary database, and its source code reveals the structure of this database. However, we plan to negotiate access to the Dimensions database, which is free for scientific use.

The minimum level of openness that can be currently achieved is making publicly available a version of MALBA that analyses pre-existing networks. We hope to achieve more but this depends on negotiations with database owners.

Author contributions

Matthias Held: Conceptualization, Data curation, Methodology, Investigation, Data analysis, Writing

Bastian Steudel: Methodology, Software, Visualization, Writing

Jochen Gläser: Funding acquisition, Conceptualization, Methodology, Project administration, Supervision, Writing

References

Chen, C. (2018). Cascading citation expansion. *Journal of Information Science Theory and Practice*, *6*(2), 6–23.

Chen, Q., Wu, T.-T., & Fang, M. (2013). Detecting local community structures in complex networks based on local degree central nodes. *Physica A: Statistical Mechanics and Its Applications*, *392*(3), 529–537. https://doi.org/10.1016/j.physa.2012.09.012

Clauset, A. (2005). Finding local community structure in networks. *Physical Review E*, *72*(2), 026132. https://doi.org/10.1103/PhysRevE.72.026132

Dilmaghani, S., Brust, M. R., Danoy, G., & Bouvry, P. (2021). Community Detection in Complex Networks: A Survey on Local Approaches. In N. T. Nguyen, S. Chittayasothorn, D. Niyato, & B. Trawiński (Eds.), *Intelligent Information and*

Database Systems (Vol. 12672, pp. 757–767). Springer International Publishing. https://doi.org/10.1007/978-3-030-73280-6_60

Fanrong, M., Mu, Z., Yong, Z., & Ranran, Z. (2014). Local Community Detection in Complex Networks Based on Maximum Cliques Extension. *Mathematical Problems in Engineering*, *2014*, 1–12. https://doi.org/10.1155/2014/653670

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, *486*(3), 75–174. https://doi.org/10.1016/j.physrep.2009.11.002

Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, *659*, 1–44. https://doi.org/10.1016/j.physrep.2016.09.002

Gläser, J. (2019). How can governance change research content? Linking science policy studies to the sociology of science. *Handbook on Science and Public Policy*, 419–447.

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

Hamann, M., Röhrs, E., & Wagner, D. (2017). Local Community Detection Based on Small Cliques. *Algorithms*, *10*(3), Article 3. https://doi.org/10.3390/a10030090

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. (2022). Know thy tools! Limits of popular algorithms used for topic reconstruction. *Quantitative Science Studies*, *3*(4), 1054–1078. https://doi.org/10.1162/qss_a_00217

Held, M., Laudel, G., & Gläser, J. (2021). Challenges to the validity of topic reconstruction. *Scientometrics*, *126*(5), 4511–4536. https://doi.org/10.1007/s11192-021-03920-3

Huang, J., Sun, H., Liu, Y., Song, Q., & Weninger, T. (2011). Towards Online Multiresolution Community Detection in Large-Scale Networks. *PLoS ONE*, *6*(8), e23829. https://doi.org/10.1371/journal.pone.0023829

Kuhn, T. (1970 [1962]). *The Structure of Scientific Revolutions*. The University of Chicago Press.

Whitley, R. (2000). *The Intellectual and Social Organization of the Sciences*. Clarendon Press.