RESEARCH ARTICLE

Enhancing direct citations: A comparison of relatedness measures for community detection in a large set of PubMed publications

Per Ahlgren[1], Yunwei Chen[2], Cristian Colliander[3,4], and Nees Jan van Eck[5]

[1]Department of Statistics, Uppsala University, Uppsala (Sweden)
[2]Scientometrics & Evaluation Research Center (SERC), Chengdu Library and Information Center of Chinese Academy of Sciences, Chengdu, 610041 (China)
[3]Department of Sociology, Inforsk, Umeå University, Umeå (Sweden)
[4]University Library, Umeå University, Umeå (Sweden)
[5]Centre for Science and Technology Studies, Leiden University (The Netherlands)

Keywords: citation-based relatedness measures, clustering solution accuracy, community detection, enhancing direct citations, MeSH, text-based relatedness measures

ABSTRACT

The effects of enhancing direct citations, with respect to publication–publication relatedness measurement, by indirect citation relations (bibliographic coupling, cocitation, and extended direct citations) and text relations on clustering solution accuracy are analyzed. For comparison, we include each approach that is involved in the enhancement of direct citations. In total, we investigate the relative performance of seven approaches. To evaluate the approaches we use a methodology proposed by earlier research. However, the evaluation criterion used is based on MeSH, one of the most sophisticated publication-level classification schemes available. We also introduce an approach, based on interpolated accuracy values, by which overall relative clustering solution accuracy can be studied. The results show that the cocitation approach has the worst performance, and that the direct citations approach is outperformed by the other five investigated approaches. The extended direct citations approach has the best performance, followed by an approach in which direct citations are enhanced by the BM25 textual relatedness measure. An approach that combines direct citations with bibliographic coupling and cocitation performs slightly better than the bibliographic coupling approach, which in turn has a better performance than the BM25 approach.

1. INTRODUCTION

Community detection in citation networks, which is the topic of this paper, can be performed in order to analyze both the obvious and the more subtle relations between scientific publications, as well as the identification of subfields of science (e.g., Chen & Redner, 2010; Klavans & Boyack, 2017; Waltman & Van Eck, 2012). In the context of networks, communities are clusters of closely connected nodes within a network. Communities of this kind are found not only in citation networks, but also in many other networks, such as biological networks, the World Wide Web, social networks, and collaboration works (Girvan & Newman, 2002).

Citation networks originate from the relationships between citing and cited publications. Community structure can often be observed in these networks, because publications dealing

an open access journal

Check for updates

Citation: Ahlgren, P., Chen, Y., Colliander, C., & van Eck, N. J. (2020). Enhancing direct citations: A comparison of relatedness measures for community detection in a large set of PubMed publications. *Quantitative Science Studies*. *1*(2), 714–729. https://doi.org/10.1162/qss_a_00027

DOI:
https://doi.org/10.1162/qss_a_00027

Received: 21 August 2019
Accepted: 27 January 2020

Corresponding Author:
Per Ahlgren
per.ahlgren@uadm.uu.se

Handling Editor:
Vincent Larivière

The MIT Press

with a given topic tend to cite similar publications with respect to topic. Communities in a citation network thereby contain similar publications regarding a single topic or a set of related topics. For a given field, community detection in a citation network can be used to uncover related publications. The detected subfields, and interrelations between them, might then be useful for researchers and policy makers, because the subfields and their interrelations indicate the whole pattern of the field at a glance.

Although several studies on community detection in citation networks have been performed in recent years, we have not found many such studies that discriminate, based on some notion of importance, between citation relations. However, Small (1997) explored the idea of combining direct citation information with indirect citation information. Persson (2010) used weighted direct citations, where the citations were weighted by shared references and cocitations in order to decompose a citation network. Persson investigated the field of library and information science and obtained meaningful subfields by removing direct citations with weights below a certain threshold and by removal of less frequently cited publications. The study by Fujita, Kajikawa, et al. (2014) constitutes another example of a study using weighted direct citations. Different types of weighted citation networks were studied with regard to detection of emerging research fields, where the weights were based, for instance, on reference lists and keyword similarity. Chen, Fengxia, and Wang (2013) proposed a community discovery algorithm to uncover semantic communities in a citation semantic link network. In that study, direct citations were weighted on the basis of common keywords. A fifth example of a study that discriminates between direct citation relations is the work by Chen, Xiao, Deng, and Zhang (2017). These authors used two publication data sets and modularity-based clustering of publications, and compared clustering solutions obtained on the basis of four approaches, where the main difference between these approaches is how the relatedness of two publications is defined. One of the approaches is based on direct citations, whereas the other three weight the direct citations in three different ways. All of the latter three approaches use textual similarities as weights, and two of them take term position information into account. The study by Chen et al. (2017) inspired us to perform another study, in which we investigated the relative clustering solution accuracy of nine publication–publication relatedness measures (Ahlgren, Chen, et al., 2019).

One can distinguish between two types of methods used for citation network community detection. One type consists of methods based only on the topological structure of the network, that is, the arrangement of publications (nodes) and citation relations (links) (e.g., Boyack & Klavans, 2014; Chen & Redner, 2010; Haunschild, Schier, et al., 2018; Kajikawa, Yoshikawa, et al., 2008; Klavans & Boyack, 2017; Kusumastuti, Derks, et al., 2016; Ruiz-Castillo & Waltman, 2015; Sjögårde & Ahlgren, 2018, 2020; Subelj, Van Eck, & Waltman, 2016; Waltman & Van Eck, 2012; Yudhoatmojo & Samuar, 2017), whereas the other type consists of methods that also use publication content, represented by text. To take both topological structure and content into account in an analysis of citation networks might be fruitful. This has been done, as we have seen, in community detection analyses and with regard to direct citations (Chen, Fengxia, & Wang, 2013; Chen et al., 2017; Fujita et al., 2014), but it has also been done in studies in which bibliographic coupling or cocitation have been used as citation relations (e.g., Ahlgren & Colliander, 2009; Glänzel & Thijs, 2017; Meyer-Brötz, Schiebel, & Brecht, 2017; Yu, Wang, et al., 2017). However, taking both topological structure and content into account has also been done in studies not involving community detection. Cohn and Hofmann (2001) described a joint probabilistic model for modeling the contents and interconnectivity of publication collections such as sets of research publications, and Hamedani, Kim, and Kin (2016) presented a novel method called SimCC that considers both citations and content in the calculation of publication–publication similarity.

Even if the last two papers referred to in the preceding paragraph did not involve community detection in citation networks, they provide ideas that can be used for community detection in such networks. Indeed, in this study we use both topological structure and content information in citation networks to detect communities. We build on the earlier work by Chen et al. (2017) on the weighting of citation relations, as well as on the work by Waltman, Boyack, et al. (2017, 2019) on a principled methodology for evaluating the accuracy of clustering solutions using different relatedness measures. In this study, which is an extension of the study performed by Ahlgren et al. (2019), the effects of enhancing direct citations, with respect to publication–publication relatedness measurement, by indirect citation relations and text relations on clustering accuracy are analyzed. In total, we investigate seven approaches, compared to six in Ahlgren et al. (2019). In one of these, direct citations are enhanced by both bibliographic coupling and cocitation, whereas text relations are used to enhance direct citations in another approach. We also include an indirect citation relations enhancing approach that takes direct citation relations within an extended set of publications into account. We include in the study, for comparison reasons, each approach that is involved in the enhancement of direct citations. We also introduce a methodology by which overall relative clustering solution accuracy can be studied. This methodology was not used in Ahlgren et al. (2019).

Compared to the study by Chen et al. (2017), a considerably larger publication set is used in our study, as well as a more sophisticated evaluation methodology, in which an external subject classification scheme, Medical Subject Headings (MeSH), is used. MeSH is one of the most sophisticated publication-level classification schemes available. Moreover, in contrast to the earlier work, we use a different approach regarding the combination of direct citations and text relations. Compared to Waltman et al. (2017, 2019), these authors did not evaluate hybrid relatedness approaches (approaches combining citation and text relations). Further, citation-only approaches were only compared to other such approaches in their analysis, and the same was the case for text-only approaches. An advantage of our study, however, is that comparisons across such approach groups could be made due to the use of MeSH as an independent evaluation criterion.

The remainder of the paper is organized as follows. In the next section, we deal with data and methods, whereas the results of the study are reported in the third section. In the final section, we provide a discussion as well as conclusions.

2. DATA AND METHODS

Because direct citations are used in the study, we needed a sufficiently long publication period. We decided to use a five-year period, namely 2013–2017. Initially, a set of 4,260,452 MEDLINE—the largest subset of PubMed—publications were retrieved from PubMed, where the query included a reference to the publication period. The following query was used: *MEDLINE[SB] AND ("2013/01/01"[PDat] : "2017/12/31"[PDat])*. From the initially retrieved set, we filtered out those publications with a print year in the interval 2013–2017, which yielded a set of 4,191,763 publications. Because PubMed does not contain citation relations between publications, we also use Web of Science (WoS) data. The next step was then to match, using PMID data, each publication in this set of publications to publications included in the in-house version of the WoS database available at the Centre for Science and Technology Studies (CWTS) at Leiden University, which yielded a set of 3,577,358 publications. From this latter set, we selected each publication p such that p satisfies each of the following four conditions:

1. p has a WoS publication year in the period 2013–2017.
2. p is of WoS document type *Article* or *Review*.

3. p has both an abstract and a title with respect to its WoS record.
4. p has a citation relation to at least one publication p' such that p' satisfies points 1–3 in this list.

A total of 2,941,119 publications satisfied all four conditions. However, 10 of these publications were removed, because they are not indexed with MeSH descriptors in PubMed. Such descriptors are needed by our evaluation methodology (see subsection 2.3). Our final publication set, P_{MEDLINE}, then consists of 2,941,109 publications.

2.1. Investigated approaches

As stated above, we compare seven approaches to publication community detection in this study. The main difference between the approaches is how the relatedness of two publications is defined. Five of the approaches—DC (direct citations), EDC (extended direct citations), BC (bibliographic coupling), CC (cocitation), and DC-BC-CC (combination of direct citations, bibliographic coupling, and cocitation)—use only citation relations. Of the remaining two approaches, BM25 and DC-BM25, BM25 uses only text relations, whereas DC-BM25 combines direct citations with text relations. We now describe the seven approaches in more detail.

DC

In DC, the relatedness of two publications i and j, r_{ij}^{DC}, is defined as

$$r_{ij}^{\text{DC}} = \max\left(c_{ij}, c_{ji}\right) \tag{1}$$

where c_{ij} is 1 if i cites j, 0 otherwise. Thus, the relatedness is 1 if there is a direct citation from i to j or such a relation from j to i, otherwise the relatedness is 0.

EDC

The basic idea of this approach, in which direct citations are enhanced by indirect citation relations, is to take into account not only direct citation relations within the set of publications under consideration, in our case P_{MEDLINE}, but also direct citation relations within an extended set of publications. Let N be the number of publications under consideration, the so-called *focal* publications in the terminology of Waltman et al. (2017, 2019). In order to cluster the focal publications 1, …, N, we also take the publications $N + 1$, …, N^{EXT} into account, where each j ($j = N + 1$, …, N^{EXT}) has a direct citation relation with at least two of the focal publications. The relatedness of i and j, r_{ij}^{EDC}, where $i = 1$, …, N and $j = 1$, …, N^{EXT}, is defined as

$$r_{ij}^{\text{EDC}} = \max\left(c_{ij}, c_{ji}\right) \tag{2}$$

where c_{ij} and c_{ji} are as in Eq. 1. Thus, the same relatedness measure is used in the EDC approach as in the DC approach. However, the former approach also considers direct citation relations between the focal publications and the additional $N^{\text{EXT}} - N$ publications. Note that direct citation relations are not considered within the additional publications (i takes values in the set {1, …, N}. In this study, $N^{\text{EXT}} - N = 7{,}899{,}313$, and the additional publications are published in the period 1980–2012. Thus, because the focal publications are published in the period 2013–2017, each additional publication is cited by at least two focal publications.

BC

Here, the relatedness of i and j, r_{ij}^{BC}, is defined as the number of shared cited references in i and j, where only cited references pointing to publications covered by the CWTS in-house version of WoS are taken into account.

CC

The relatedness of i and j, r_{ij}^{CC}, is defined as the number of publications that cite both i and j.

BM25

The first step in this approach is to identify terms in the titles and abstracts of the publications in P_{MEDLINE}. Here a *term* is defined as a noun phrase: a sequence s of words of length n ($n \geq 1$) such that (a) each word in s is either a noun or an adjective, and (b) s ends with a noun. The part-of-speech tagging algorithm provided by the Apache OpenNLP 1.5.2 library is used to identify the nouns and adjectives. Plural and singular noun phrases are regarded as the same term, and shorter terms appearing in longer terms are not counted.

The BM25 approach involves the BM25 measure, a well-known query-publication similarity measure in information retrieval research (Sparck Jones, Walker, & Robertson, 2000a, 2000b) and, according to experimental results obtained by Boyack et al. (2011), one of the most accurate text-based measures for clustering publications. Let N be the number of publications under consideration (in our case, N is equal to $|P_{\text{MEDLINE}}| = 2{,}941{,}109$) and m the number of unique terms occurring in the N publications. Let o_{il} be the number occurrences of term l in publication i, and n_l the number of publications in which term l occurs. Further, $I(o_{il} > 0) = 1$ if $o_{il} > 0$ and 0 otherwise. The relatedness of i and j, r_{ij}^{BM25}, is then defined as

$$r_{ij}^{\text{BM25}} = \sum_{l=1}^{m} I(o_{il} > 0)\ \text{IDF}_l \frac{o_{jl}(k_1 + 1)}{o_{jl} + k_1\left(1 - b + b\frac{d_j}{\bar{d}}\right)} \tag{3}$$

where

$$\text{IDF}_l = \log \frac{N - n_l + 0.5}{n_l + 0.5} \tag{4}$$

and

$$d_j = \sum_{p=1}^{m} o_{jp}; \bar{d} = \frac{1}{N}\sum_{q=1}^{N}\sum_{p=1}^{m} o_{qp} \tag{5}$$

IDF_l is the inverse document frequency of term l, d_j the length of publication j, and $\bar{d}$ the mean length of the N publications. k_1 and b are parameters with respect to term frequency saturation and publication length normalization, respectively. For the values of these, we followed Boyack et al. (2011) and Waltman et al. (2017, 2019), and thereby used 2 and 0.75 for k_1 and b, respectively. Note that it is possible that $r_{ij}^{\text{BM25}} \neq r_{ji}^{\text{BM25}}$, that is, the BM25 measure is not symmetrical. It follows from Eq. 3 that $r_{ij}^{\text{BM25}} > 0$ if and only if there is at least one term occurring in both i and j.

DC-BC-CC

In this approach, as in EDC, direct citations are enhanced by indirect citation relations. More precisely, direct citations are enhanced by the citation relations corresponding to the approaches BC and CC. We define the relatedness of i and j, $r_{ij}^{\text{DC-BC-CC}}$, as

$$r_{ij}^{\text{DC-BC-CC}} = \alpha r_{ij}^{\text{DC}} + r_{ij}^{\text{BC}} + r_{ij}^{\text{CC}} \tag{6}$$

where α is a weight of direct citations relative to BC and CC. With this weight, one has the possibility to boost direct citations, which might be considered as stronger signals of the relatedness of two publications compared to a bibliographic coupling or a cocitation relation (Waltman & van Eck, 2012). In our analysis, we use 1 and 5 as values of α, in agreement with Waltman et al. (2017, 2019). Note, in contrast to DC and EDC, that the relatedness value of i and j in DC-BC-CC (and in DC-BM25, see below) can be positive without a direct citation between i and j.

DC-BM25

In this approach, direct citations are enhanced by text relations. We define the relatedness of i and j, $r_{ij}^{\text{DC-BM25}}$, as

$$r_{ij}^{\text{DC-BM25}} = \alpha r_{ij}^{\text{DC}} + r_{ij}^{\text{BM25}} \tag{7}$$

where α is a weight of direct citations relative to BM25. We obtain values of α in the following way. The average across all BM25 relatedness values greater than 0 is calculated, an average that turned out to be equal to 50. By setting α to 50, the DC values are put on the same scale as the BM25 relatedness values, in an average sense. By setting α to 25 (100), less (more) emphasis would be put on DC. We use all these three α values in our analysis.

When calculating r_{ij}^{X}, $X \in \{\text{BC, CC, BM25, DC-BC-CC, DC-BM25}\}$, we only consider the k-nearest neighbors to i (i.e., the k publications with the highest relatedness values with i). If j is not among the k publications with the highest relatedness values with i, $r_{ij}^{X} = 0$. Here, k is set to 20. For a sensitivity analysis, we refer the reader to Waltman et al. (2019). We apply the k-nearest neighbors technique for efficiency reasons. However, we do not apply this technique in DC or EDC, because computer memory requirements are relatively modest for these two approaches.

In contrast to DC, we do not enhance EDC by BC and CC. The reason for this is that BC and CC are both indirectly taken into account in the EDC approach due to the requirement for inclusion among the focal publications. To see this, consider a publication p that meets the requirement to be added to the extended set of publications (i.e., p has a direct citation relation with at least two of the focal publications). Now, because, in our case, p is published before year 2013 (the start publication year in our study), p is cited by at least two focal publications, and thereby p gives rise to a bibliographic coupling relation between at least two focal publications. If p had been published after year 2017 (which, however, is not the case in the study), p would cite at least two focal publications, and thereby give rise to a cocitation relation between at least two focal publications.

2.2. Normalization of the relatedness measures and clustering of publications

For all seven approaches, the corresponding relatedness measures are normalized. The *normalized relatedness* of publication i with publication j is the relatedness of i with j, divided

by the total relatedness of i with all other publications that are considered. Now, without normalization, clustering solutions obtained using different relatedness measures, but associated with the same value of the resolution parameter of the clustering (see below in this section), might be far from satisfying the requirement that, with regard to accuracy, the compared solutions should have the same granularity, where the *granularity* of a solution is defined as the number of publications divided by the sum of the squared cluster sizes (Waltman et al., 2017, 2019). With the indicated normalization, the granularity requirement can be assumed to be approximately satisfied by the solutions. However, to further deal with the granularity issue, granularity–accuracy plots (GA plots) are used in the study (Waltman et al., 2017, 2019). GA plots are described in the section on evaluation of approach performance below.

In this study, we use the Leiden algorithm (Traag, Waltman, & Van Eck, 2018, 2019) to generate a series of clustering solutions for each of the relatedness measures. The Leiden algorithm is used to maximize the Constant Potts Model as quality function (Traag, Van Dooren, & Nesterov, 2011; Waltman & Van Eck, 2012). However, in EDC, an adjusted quality function is used in order to accommodate the nonfocal publications $N + 1, \ldots, N^{\text{EXT}}$ (Waltman et al., 2019). After maximization of the adjusted quality function, the cluster assignments of the nonfocal publications are disregarded, because we are only interested in the cluster assignments of the focal publications (i.e., the publications in P_{MEDLINE}). Using different values of the resolution parameter γ (0.000001, 0.000002, 0.000005, 0.00001, 0.00002, 0.00005, 0.0001, 0.0002, 0.0005, 0.001, 0.002), we obtain 11 clustering solutions for each relatedness measure. Compared to our earlier study (Ahlgren et al., 2019), we exclude the clustering solutions for the two largest resolution values used in that study (0.005 and 0.01). These clustering solutions have around 300,000 and 500,000 clusters, respectively, and most of the clusters consist of fewer than 10 publications. From a practical point of view, the utility of these detailed cluster solutions can be questioned, and we believe it makes sense to exclude them.

The normalization of the relatedness measures transforms these measures to nonsymmetrical counterparts. However, the clustering methodology we use requires that the relatedness values are symmetrical. We solve this issue in the following way. Let $\hat{r}_{ij}^{X}$ denote the relatedness of i with j with respect to approach $X \in$ {DC, EDC, BC, CC, BM25, DC-BC-CC, DC-BM25} after normalization of r_{ij}^{X}. The relatedness value for i and j given as input to the clustering algorithm is $\hat{r}_{ij}^{X} + \hat{r}_{ji}^{X}$ (i.e., the sum of the two normalized relatedness values). Clearly, then, the relatedness values are made symmetrical before being given as input to the clustering algorithm.

2.3. Evaluation of approach performance

For the evaluation of the performance of the seven approaches, an external and independent subject classification scheme, MeSH, is used. MeSH descriptors and subheadings are used to index publications in PubMed. MeSH contains more than 28,000 descriptors that are arranged hierarchically by subject categories, with more-specific descriptors arranged beneath broader descriptors (U.S. National Library of Medicine, 2019a). MeSH descriptors can be designated as major, indicating that they correspond to the major topics of the publication, whereas nonmajor descriptors are added to reflect additional topics substantively discussed within the publication. Further, approximately 80 subheadings (or qualifiers) can be used by the indexer to qualify a descriptor. Subheadings are thus not standalone terms and are only used in conjunction with a descriptor to describe specific aspects of the descriptor that are pertinent to the publication. For example, the descriptor "Ectopia Lentis" can be combined with the subheading "surgery" to specify that the publication deals with surgical treatment of the displacement of the eye's crystalline lens. Descriptors will usually be indexed with one or more subheadings.

The assignment of MeSH descriptors and subheadings to publications is based on a manual reading of these publications by human indexers (U.S. National Library of Medicine, 2019b). Relatedness measurement based on MeSH, described below, thus differs substantially from the seven evaluated relatedness approaches, as the latter are based on directly observable features in the publications (words and references), whereas assigned MeSH descriptors and subheadings are the result of a human intellectual indexing process, whose aim is to produce standardized subject descriptions.

Relatedness measurement based on MeSH is done as follows. We first calculate a weight (information content, IC) for each descriptor (Colliander & Ahlgren, 2019; Zhu, Zeng, & Mamitsuka, 2009). Let $freq(desc_i)$ denote the frequency of descriptor i (here calculated over all MEDLINE publications published within the period 2013–2017). Then

$$IC(desc_i) = -\log(P(desc_i)) \tag{8}$$

where

$$P(desc_i) = \frac{freq(desc_i) + \sum_{d \in descendants(desc_i)} freq(d)}{\sum_{k=1}^{s} \left(freq(desc_k) + \sum_{d \in descendants(desc_k)} freq(d) \right)} \tag{9}$$

where $descendants(desc_i)$ is the set of descriptors that are children, direct or indirect, to descriptor i in the MeSH tree.

We then represent each publication by a vector of length $s + (s \times m)$, where s and m are the total number of unique MeSH descriptors and the total number of unique[1] subheadings in the data set, respectively. The vector position for the ith descriptor is given by $(m + 1) \times i - m$ and the corresponding *weight for publication* l ($\omega_i(l)$) is defined as

$$\omega_i(l) = \begin{cases} 0 \text{ if } desc_i \text{ is absent in } l \\ IC(desc_i) \times 1 \text{ if } desc_i \text{ is a minor descriptor in } l \\ IC(desc_i) \times 2 \text{ if } desc_i \text{ is a major descriptor in } l \end{cases} \tag{10}$$

The vector position for the jth subheading connected to the ith descriptor is given by $(m + 1) \times i - m + j$ and the corresponding *weight for publication* l ($\phi_{ji}(l)$) is defined as

$$\phi_{ji}(l) \begin{cases} 1 \text{ if subheading } j \text{ and descriptor } i \text{ are present in } l \\ 0 \text{ otherwise} \end{cases} \tag{11}$$

Note that many descriptor–subheading pairs are nonsensical and will never exist in practice, and the subheading in such a pair will thus always take on the value 0 in the vectors.

We estimate the relatedness between the publications by the cosine similarity (Salton & McGill, 1983) between their corresponding vectors as defined above. As in the case of calculating relatedness in BC, CC, BM25, DC-BC-CC, and DC-BM25, and for the same reason, we apply the k-nearest neighbors technique. As in these five approaches, k is set to 20. We then normalize the cosine similarities in the same way as we normalize the relatedness measures of all seven approaches, resulting in $\hat{r}_{ij}^{\text{MeSH}}$. Finally, the publications in P_{MEDLINE} are clustered based on the normalized cosine similarities using the same clustering methodology, and the same set of values of the resolution parameter, as for the seven approaches.

[1] A group of MeSH descriptors that are routinely added to most articles, "check tags," are concepts of potential interest, regardless of the general subject content of the article (examples are "Human" and "Adult"). We do not include such check tags in any calculations.

The accuracy of the lth ($1 \leq l \leq 11$) clustering solution for $X \in$ {DC, EDC, BC, CC, BM25, DC-BC-CC, DC-BM25, MeSH}, where the accuracy is based on MeSH cosine similarity, symbolically $A^{X_l|\text{MeSH}}$, is defined as follows (Waltman et al., 2017, 2019):

$$A^{X_l|\text{MeSH}} = \frac{1}{N}\sum_{i,j} I\left(c_i^{X_l} = c_j^{X_l}\right)\hat{r}_{ij}^{\text{MeSH}} \tag{12}$$

where $i, j \in P_{\text{MEDLINE}}$, $c_i^{X_l}$ ($c_j^{X_l}$) is a positive integer denoting the cluster to which publication i (j) belongs with respect to the lth clustering solution for X, $I(c_i^{X_l} = c_j^{X_l})$ is 1 if its condition is true, otherwise 0, and $\hat{r}_{ij}^{\text{MeSH}}$ the normalized MeSH cosine similarity of i with j. Recall that DC-BC-CC (DC-BM25) has two (three) variants, $\alpha \in \{1, 5\}$ ($\alpha \in \{25, 50, 100\}$), and that we thereby, in total, work with 11 relatedness measures. Note that we want to compare, with respect to clustering solution accuracy, the 10 measures distinct from MeSH. However, we also include clustering solutions based on the MeSH cosine similarity in a part of the evaluation exercise (cf. Section 3.1). The accuracy results obtained for MeSH give an upper bound for the results that can be obtained when the relatedness measures of the seven approaches are used to cluster the publications and accuracy is based on MeSH cosine similarity. We remind the reader that the value of the resolution parameter γ is held constant across the seven approaches and MeSH regarding the kth clustering solution.

We visualize the evaluation results by using GA plots. The use of such plots is a way to counteract the difficulty that the requirement that, with regard to accuracy, the compared clustering solutions should have the same granularity is only approximately satisfied. In a GA plot, the horizontal axis represents granularity (as defined above), whereas the vertical axis represents accuracy. For a given approach, such as DC, a point in the plot represents the accuracy and granularity of a clustering solution, obtained using a certain resolution value of γ. Further, a line is connecting the points of the approach, where accuracy values for granularity values between points are estimated by the technique Piecewise Cubic Hermite Interpolation. Based on the interpolations, the performance of the approaches can be compared at a given granularity level. The interpolation technique is described in the Appendix.

3. RESULTS

In this section, we first present performance results for the seven tested approaches using GA plots. We then deal with relative overall approach performance, where a summary value based on interpolated accuracy values is obtained for each of the 10 relatedness measures.

3.1. Performance results: GA plots

We present three figures containing GA plots. The first plot contains curves for DC and the other citation-based approaches, the second for DC and the text-based approaches, whereas the last plot contains curves for DC and the best performing approaches. As should be clear from section 2, MeSH is consistently used as the evaluation criterion. Note that all three plots contain a curve also for MeSH, where such a curve represents an upper bound for the performance of the seven approaches. One might ask what the meaning, in terms of number of clusters, of different granularity levels is. When the granularity is around 0.0001, a clustering solution typically has 500 significant clusters (defined as the number of clusters with 10 or more publications). When the granularity is around 0.001 (0.01), a clustering solution typically has 5,000 (50,000) significant clusters.

The GA plot of Figure 1 visualizes the accuracy results of enhancing DC by indirect citations. The performance of EDC and the combination of DC with BC and CC ($\alpha = 1, 5$), as well as the performance of DC, BC, and CC, is shown. CC exhibits the worst performance among

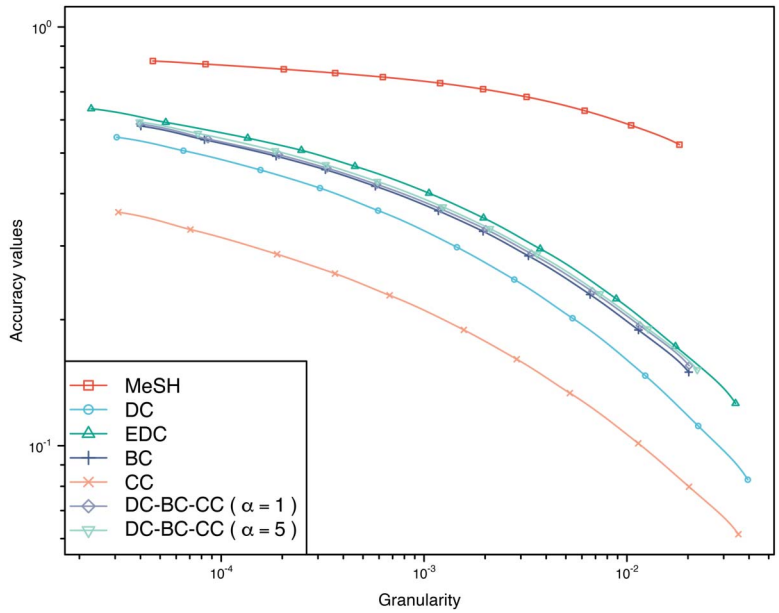

Figure 1. GA plot for comparing the approaches DC, EDC, BC, CC, and the two variants of DC-BC-CC. MeSH is used as the evaluation criterion.

the citation-based approaches. EDC has the best performance, followed by DC-BC-CC (α = 5). BC performs slightly worse than DC-BC-CC (α = 1), and DC is outperformed by all three approaches in which DC is enhanced by indirect citation relations.

In Figure 2, a GA plot that shows the results of enhancing DC by BM25, and thereby by textual relations, is given (α = 25, 50, 100). The plot also shows the performance of DC and

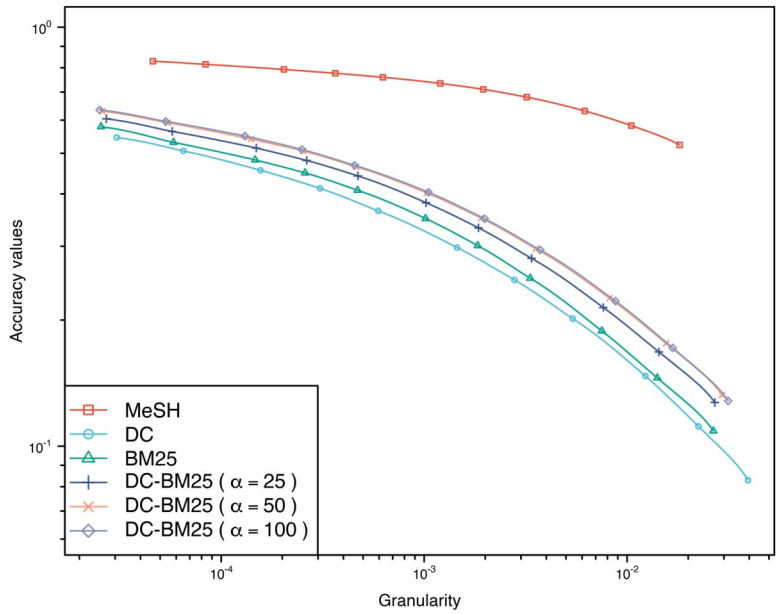

Figure 2. GA plot for comparing the approaches DC, BM25, and the three variants of DC-BM25. MeSH is used as the evaluation criterion.

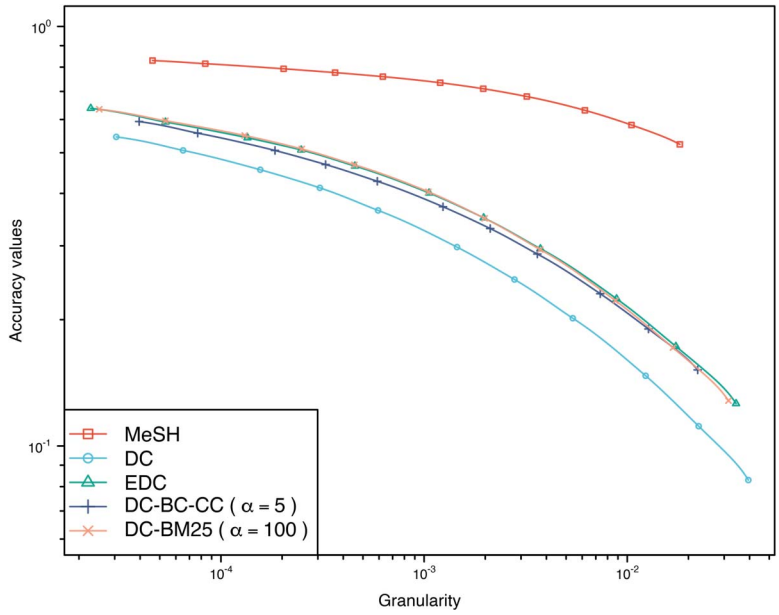

Figure 3. GA plot for comparing DC, EDC, DC-BC-CC (α = 5), and DC-BM25 (α = 100). MeSH is used as the evaluation criterion.

BM25. BM25 performs better than DC, but is outperformed by all three DC-BM25 variants. Of these, those with α equal to 50 and 100 perform about equally well, and better than the variant that puts less emphasis on DC (α = 25).

Our final GA plot (Figure 3) shows the performance of DC and the best performing approaches, namely EDC, DC-BC-CC (α = 5), and DC-BM25 (α = 100). Extended direct citations (i.e. EDC) and enhancing DC by BM25 yield the best performance. DC-BC-CC, where DC is enhanced by the combination of BC and CC, then performs worse than DC-BM25, whereas DC, as we already know (Figures 1 and 2), has the worst performance. Although the lines of EDC and DC-BM25 are for a large part overlapping in Figure 3, it seems that EDC performs slightly better than DC-BM25 for clustering solutions with a higher granularity (thus solutions with a higher number of clusters). This difference is further studied in the next subsection.

3.2. Performance results: Relative overall clustering solution accuracy

In this subsection, we complement the picture of relative performance given in the preceding subsection. We do this by introducing a methodology that results in one numerical value per relatedness measure. This value, which summarizes the relative clustering solution accuracy for the corresponding measure, is introduced as an approximate measure for easier comprehension of GA plots.

We let $p_j(x)$ denote the interpolation function for the jth ($1 \leq j \leq 10$) relatedness measure[2], where x is a granularity value and Piecewise Cubic Hermite Interpolation (see Appendix) is used. We then define the average interpolated accuracy value with respect to x, $p_{Avg}(x)$, as

$$p_{Avg}(x) = \frac{1}{m}\sum_{j=1}^{m} p_j(x) \tag{13}$$

where m, in this context, is equal to 10.

[2] We do not consider our relatedness evaluator, MeSH, in this part of the evaluation exercise.

Let a and b be the minimum and maximum values, respectively, such that for each relatedness measure j, $p_j(a)$, and $p_j(b)$ are defined (extrapolation is not used). Let $s^l = (a, \ldots, b)$ be a sequence of l evenly spaced values between a and b, and let s_i^l denote the ith value in s^l. Then a reasonable summary value for the relative clustering solution accuracy of relatedness measure j is defined as

$$acc_j = \frac{1}{l}\sum_{i=1}^{l} \frac{p_j(s_i^l)}{p_{avg}(s_i^l)} \tag{14}$$

For a given relatedness measure j, and for each value s_i^l in s^l, the interpolated accuracy value with respect to s_i^l is divided by the average interpolated accuracy value with respect to s_i^l across the relatedness measures. Then the mean across the l ratios is obtained, and constitutes the summary value for the relative clustering solution accuracy of relatedness measure j. Note that $acc_j = 1$ corresponds to average performance. In the study, l was set to 500.

The bar chart of Figure 4 visualizes the relative overall clustering solution accuracy of the 10 relatedness measures. The measures, corresponding to the bars, are ordered descending from left to right according to their accuracy values (Eq. 14). Further, the color of a bar indicates measure type. The red bar corresponds to direct citations (DC), the two blue bars to indirect citations (BC and CC), the three green bars to DC enhanced by indirect citations (the two DC-BC-CC variants and EDC), the purple bar to textual relations (BM25), and the three orange bars to DC enhanced by textual relations (the three variants of DC-BM25). The horizontal dotted line indicates average performance.

EDC has the highest overall performance, an outcome that provides additional information compared to the GA plot of Figure 3. Similarly, from the point of view of overall performance, DC-BM25 (α = 100) performs better than DC-BM25 (α = 50) (cf. the GA plot of Figure 2). The overall performance order of the two DC-BC-CC variants and BC agrees with the GA plot of Figure 1, and the overall performance order of DC, CC, and BM25 agrees with the GA plots of

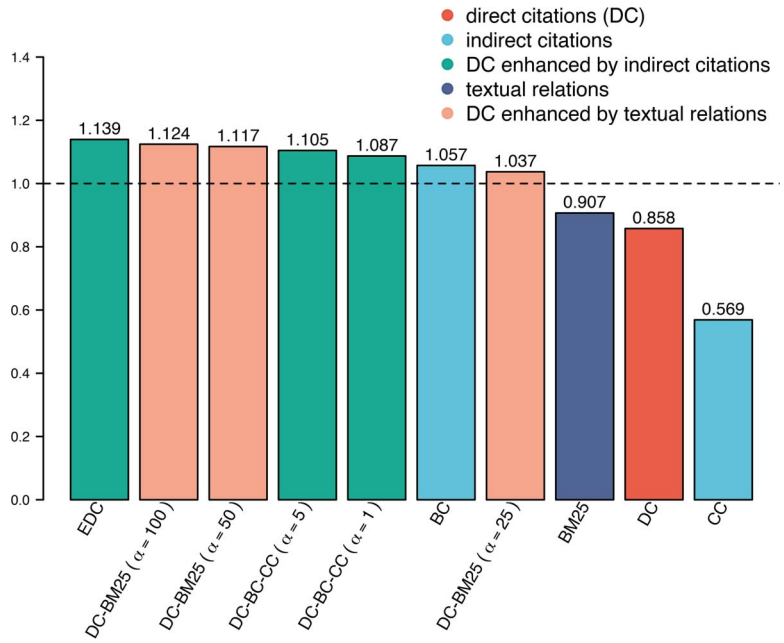

Figure 4. Relative overall clustering solution accuracy for the 10 relatedness measures according to Eq. 14.

Figures 1 and 2. In general, then, our conclusions based on the relative clustering solution accuracy values are in line with the conclusions that can be drawn based on the GA plots.

4. DISCUSSION AND CONCLUSIONS

We have analyzed the effects of enhancing direct citations, with respect to publication–publication relatedness measurement, by indirect citation relations and text relations on clustering solution accuracy. We used an approach based on MeSH, one of the most sophisticated publication-level classification schemes available, as the independent evaluation criterion. Seven approaches were investigated, and the results show that using extended direct citations (EDC), as well as enhancing direct citations (DC) with bibliographic coupling (BC) and cocitation (CC) or text relations (BM25), gives rise to substantial performance gains relative to DC. The best performance was obtained by EDC, followed by DC-BM25 and DC-BC-CC. Thus, in our analysis, extended direct citations give the best performance and, interestingly, enhancing direct citations by text relations gives rise to better performance compared to enhancing direct citations by bibliographic coupling and cocitation.

The poor performance of CC has been observed in earlier research (Klavans & Boyack, 2017; Waltman et al., 2017, 2019) and was expected. Clearly, a publication that has not received any citations is not cocited with another publication, and can therefore not be adequately clustered. In the study by Klavans and Boyack (2017), in which a more expansive EDC variant was used compared to our variant, EDC yielded more accurate clusters than BC. In this respect, our study reinforces the results of Klavans and Boyack (2017).

Waltman et al. (2017, 2019) compared DC, EDC, BC, CC, and DC-BC-CC ($\alpha = 1, 5$), using BM25 as the evaluation criterion and a considerably smaller publication set than the publication set of our analysis. Our results for these citation-based approaches demonstrate the same pattern as the results of these authors. This supports the robustness of the results for the five citation-based approaches, because the two studies used different publication sets and different evaluation criteria.

In our study, BM25 is outperformed by EDC. Boyack and Klavans (2018), though, concluded that clusters that were obtained on the basis of the text-only relatedness measures used in their study are as accurate as those that were obtained on the basis of EDC. However, a different evaluation criterion, compared to ours, was used in the study.

Chen et al. (2017) used the TF-IDF term weighting approach combined with the cosine similarity measure in order to weight direct citations by textual similarities. We tested the same approach (without taking term position information into account), as well as an approach in which BM25 is used for the weighting of direct citations. These two approaches, called *DC-TF-IDF and DC-BM25* (weighted links), were outperformed, though, by DC-BM25, DC-BC-CC and BC. Note that, for DC-TF-IDF and DC-BM25 (weighted links), and in contrast to DC-BM25, a necessary (but not sufficient) condition for obtaining a positive relatedness value for two publications *i* and *j* is that there is a direct citation from *i* to *j*, or conversely.

A limitation of our study is that it could be argued that the MeSH approach is not fully independent of relatedness measures based on text in abstracts and titles of publications, because the indexers who assign MeSH terms to publications partially rely on the title and full text of publications. Therefore, the MeSH approach might not be fully independent of the BM25 and DC-BM25 approaches. However, MeSH constitutes a controlled vocabulary, whereas BM25 makes use of an uncontrolled vocabulary, the source of which is the authors of the publications. In view of this, we believe that the MeSH approach is sufficiently different from approaches that make use of terms in abstracts and titles.

For an enhancement of EDC by BM25, which intuitively is reasonable, we obtained corresponding results in the study. These showed that EDC-BM25 performed almost as well as the best performing approach (EDC). However, for efficiency reasons, we had to use a methodology that deviates from that used in EDC. Due to demanding computer memory requirements, we needed to apply the *k*-nearest neighbor technique in the case of EDC-BM25. This was not needed in the case of EDC. We suspect that this is the reason behind the somewhat counterintuitive result that EDC-BM25 did not outperform the other approaches.

Finally, as it does not follow that two clustering solutions with similar accuracy also have similar groupings of publications into clusters, in future studies we aim to further compare the clustering solutions to deepen the insight into how solutions based on different relatedness measures diverge.

ACKNOWLEDGMENTS

We would like to thank two anonymous reviewers for their valuable comments on an earlier version of this paper.

AUTHOR CONTRIBUTIONS

Per Ahlgren: Conceptualization, Methodology, Formal analysis, Writing—original draft, Writing—review & editing. Yunwei Chen: Conceptualization, Methodology, Writing—original draft, Writing—review & editing. Cristian Colliander: Conceptualization, Methodology, Software, Formal analysis, Writing—original draft, Writing—review & editing, visualization. Nees Jan van Eck: Conceptualization, Methodology, Software, Formal analysis, Writing—original draft, Writing—review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

The article processing charge (APC) is covered by the National Key Research and Development Program of China (Grant No. 2017YFB1402400).

DATA AVAILABILITY

The data used in this paper were partly 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 (https://clarivate.com/products/web-of-science).

REFERENCES

Ahlgren, P., & Colliander, C. (2009). Document-document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics*, *3*(1), 49–63. https://doi.org/10.1016/j.joi.2008.11.003

Ahlgren, P., Chen Y. W., Colliander, C., & Van Eck, N. J. (2019). Community detection using citation relations and textual similarities in a large set of PubMed publications. Accepted for publication in *Proceedings of the 17th International Conference on Scientometrics and Informetrics*.

Boyack, K. W., & Klavans, R. (2014). Including cited non-source items in a large-scale map of science: What difference does it make? *Journal of Informetrics*, *8*(3), 569–580. https://doi.org/10.1016/j.joi.2014.04.001

Boyack, K. W., & Klavans, R. (2018). Accurately identifying topics using text: Mapping PubMed. *Proceedings of the 23rd International Conference on Science and Technology Indicators—STI 2018*, 107–115.

Boyack, K. W., Newman, D., Duhon, R. J., Klavans, R., Patek, M., Biberstine, J. R., Schijvenaars, B., Skupin, A., Ma, N., & Börner, K. (2011). Cluster-ing more than two million biomedical publications: Comparing the accuracies of nine text-based similarity approaches. *PLOS ONE*, *6*(3), e18029. https://doi.org/10.1371/journal.pone.0018029

Chen, P., & Redner, S. (2010). Community structure of the physical review citation network. *Journal of Informetrics*, *4*(3), 278–290. https://doi.org/10.1016/j.joi.2010.01.001

Chen, W., Fengxia, Y., & Wang, Y. (2013). Community discovery algorithm of citation semantic link network. *6th International Symposium on Computational Intelligence and Design* (Vol. 2), 289–292. https://doi.org/10.1109/ISCID.2013.186

Chen Y. W., Xiao X., Deng Y., & Zhang, Z. (2017). A weighted method for citation network community detection. *Proceedings of the 16th International Conference on Scientometrics and Informetrics—ISSI 2017*, 58–67.

Cohn, D., & Hofmann, T. (2001). The missing link—A probabilistic model of document content and hypertext connectivity. In T. K. Leen et al. (Eds.), *Advances in neural information processing systems 13* (pp. 430–436). Cambridge, MA: MIT Press.

Colliander, C., & Ahlgren, P. (2019). Comparison of publication-level approaches to ex-post citation normalization. *Scientometrics*, *120*(1), 283–300. https://doi.org/10.1007/s11192-019-03121-z

Fritsch, F. N., & Butland, J. (1984). A method for constructing local monotone piecewise cubic interpolants. *Siam Journal on Scientific and Statistical Computing*, *5*(2), 300–304. https://doi.org/10.1137/0905021

Fritsch, F. N., & Carlson, R. E. (1980). Monotone piecewise cubic interpolation. *Siam Journal on Numerical Analysis*, *17*(2), 238–246. https://doi.org/10.1137/0717021

Fujita, K., Kajikawa, Y., Mori, J., & Sakata, I. (2014). Detecting research fronts using different types of weighted citation networks. *Journal of Engineering and Technology Management*, *32*, 129–146. https://doi.org/10.1016/j.jengtecman.2013.07.002

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *PNAS*, *99*(12), 7821–7826. https://doi.org/10.1073/pnas.122653799

Glänzel, W., & Thijs, B. (2017). Using hybrid methods and "core documents" for the representation of clusters and topics: the astronomy dataset. *Scientometrics*, *111*(2), 1071–1087. https://doi.org/10.1007/s11192-017-2301-6

Hamedani, M. R., Kim, S. W., & Kin, D. J. (2016). SimCC: A novel method to consider both content and citations for computing similarity of scientific papers. *Information Sciences*, *334–335*, 273–292. https://doi.org/10.1016/j.ins.2015.12.001

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

Kajikawa, Y., Yoshikawa, J., Takeda, Y., & Matsushima, K. (2008). Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy. *Technological Forecasting and Social Change*, *75*(6), 771–782. https://doi.org/10.1016/j.techfore.2007.05.005

Klavans, R., & Boyack, K. W. (2017). Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge? *Journal of the Association for Information Science and Technology*, *68*(4), 984–998. https://doi.org/10.1002/asi.23734

Kusumastuti, S., Derks, M. G., Tellier, S., Di Nucci, E., Lund, R., Mortensen, E. L., & Westendorp, R. G. (2016). Successful ageing: A study of the literature using citation network analysis. *Maturitas*, *93*, 4–12. https://doi.org/10.1016/j.maturitas.2016.04.010

Meyer-Brötz, F., Schiebel, E., & Brecht, L. (2017). Experimental evaluation of parameter settings in calculation of hybrid similarities: effects of first- and second-order similarity, edge cutting, and weighting factors. *Scientometrics*, *111*(3), 1307–1325. https://doi.org/10.1007/s11192-017-2366-2

Persson, O. (2010). Identifying research themes with weighted direct citation links. *Journal of Informetrics*, *4*(3), 415–422. https://doi.org/10.1016/j.joi.2010.03.006

Ruiz-Castillo, J., & Waltman, L. (2015). Field-normalized citation impact indicators using algorithmically constructed classification systems of science. *Journal of Informetrics*, *9*(1), 102–117. https://doi.org/10.1016/j.joi.2014.11.010

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Sjögårde, P., & Ahlgren, P. (2018). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics. *Journal of Informetrics*, *12*(1), 133–152. https://doi.org/10.1016/j.joi.2017.12.006

Sjögårde, P., & Ahlgren, P. (2020). Granularity of algorithmically constructed publication-level classifications of research publications: Identification of specialties. *Quantitative Science Studies*, *1*(1), 207–238. https://doi.org/10.1162/qss_a_00004

Small, H. (1997). Update on science mapping: Creating large document spaces. *Scientometrics*, *38*(2), 275–293. https://doi.org/10.1007/BF02457414

Sparck Jones, K., Walker, S., & Robertson, S. E. (2000a). A probabilistic model of information retrieval: Development and comparative experiments: Part 1. *Information Processing and Management*, *36*(6), 779–808. https://doi.org/10.1016/S0306-4573(00)00015-7

Sparck Jones, K., Walker, S., & Robertson, S. E. (2000b). A probabilistic model of information retrieval: Development and comparative experiments: Part 2. *Information Processing and Management*, *36*(6), 809–840. https://doi.org/10.1016/S0306-4573(00)00016-9

Subelj, L., Van Eck, N. J., & Waltman, L. (2016). Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PLOS ONE*, *11*(4), e0154404. https://doi.org/10.1371/journal.pone.0154404

Traag, V. A., Van Dooren, P., & Nesterov, Y. (2011). Narrow scope for resolution-limit-free community detection. *Physical Review E*, *84*(1), 016114. https://doi.org/10.1103/PhysRevE.84.016114

Traag, V. A., Waltman, L., & Van Eck, N. J. (2018). CWTSLeiden/networkanalysis [Source code]. Zenodo. https://doi.org/10.5281/zenodo.1466831

Traag, V. A., Waltman, L., & Van Eck, N. J. (2019). From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific Reports*, *9*, 5233. https://doi.org/10.1038/s41598-019-41695-z

U.S. National Library of Medicine. (2019a). *Introduction to MeSH*. Retrieved from https://www.nlm.nih.gov/mesh/introduction.html.

U.S. National Library of Medicine. (2019b). *The Indexing Process*. Retrieved from https://www.nlm.nih.gov/bsd/indexing/training/TIP_010.html.

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

Waltman, L., Boyack, K. W., Colavizza, G., & Van Eck, N. J. (2017). A principled methodology for comparing relatedness measures for clustering publications. In *Proceedings of the 16th International Conference on Scientometrics and Informetrics—ISSI 2017*, 691–702.

Waltman, L., Boyack, K. W., Colavizza, G., & Van Eck, N. J. (2019). A principled methodology for comparing relatedness measures for clustering publications. arXiv:1901.06815.

Yu, D. J., Wang, W. R., Zhang, S., Zhang, W. Y., & Liu, R. Y. (2017). Hybrid self-optimized clustering model based on citation links and textual features to detect research topics. *PLOS ONE*, *12*(10), e0187164. https://doi.org/10.1371/journal.pone.0187164
Yudhoatmojo, S. B., & Samuar, M. A. (2017). Community detection on citation network of DBLP data sample set using LinkRank Algorithm. *Procedia Computer Science*, *124*, 29–37. https://doi.org/10.1016/j.procs.2017.12.126
Zhu, S., Zeng, J., & Mamitsuka, H. (2009). Enhancing MEDLINE document clustering by incorporating MeSH semantic similarity. *Bioinformatics*, *25*(15), 1944–1951. https://doi.org/10.1093/bioinformatics/btp338

APPENDIX: PIECEWISE CUBIC HERMITE INTERPOLATION

In our context, we want an interpolation function that is smooth in the following sense: The function belongs to the class C^1 (i.e., it is differentiable and its derivate is continuous). Moreover, we want the interpolation function to be shape preserving. This is connected to the fact that monotonicity must be guaranteed, because, for any relatedness measure, an increase in granularity will always cause a decrease in accuracy (Waltman et al., 2017, 2019). Linear interpolation will not do (not smooth), a high-order polynomial will not do (not smooth), and "standard" spline interpolation might not do (monotonicity not guaranteed). In this study, we use Piecewise Cubic Hermite Interpolation, an interpolation technique that satisfies the first condition (membership in the class C^1) indicated above. The monotonicity condition is satisfied by our use of Eq. 17 below. We now describe the interpolation technique in question.

For a set of data points (x_i, y_i) $(i = 1, \ldots, n)$, where $x_i < x_{i+1}$ $(i = 1, \ldots, n-1)$, a piecewise interpolation function $p(x) \in C^1[x_1, x_n]$ is defined such that for $i = 1, \ldots, n$

$$\begin{aligned} p(x_i) &= y_i \\ p'(x_i) &= d_i \end{aligned} \tag{15}$$

where d_i are the approximations to the derivatives of f at x_i, and where f is the underlying, unknown function that we want to approximate with interpolation. Now, let $\Delta_i = \frac{(y_{i+1}-y_i)}{(x_{i+1}-x_i)}$ and $h_i = (x_{i+1} - x_i)$, then for each $i = 1, \ldots, n-1$

$$p(x) = y_i + d_i(x - x_i) + \frac{-2d_i - d_{i+1} + 3\Delta_i}{h_i}(x - x_i)^2 + \frac{d_i + d_{i+1} - 2\Delta_i}{h_i^2}(x - x_i)^3 \tag{16}$$

is a cubic polynomial interpolation function defined on the subinterval $[x_i, x_{i+1}]$ (e.g., Fritsch & Carlson, 1980) and is the function used in this study.

There are several ways to calculate the approximations to the derivatives, but only some approaches guarantee that $p(x)$ is monotonic in each interval. One straightforward method, which guarantees monotonicity and which we use in this study, is given by Fritsch and Butland (1984). For $i = 2, \ldots, n-1$

$$d_i = \frac{\Delta_{i-1}\Delta_i}{\alpha_i\Delta_i + (1-\alpha)\Delta_{i-1}} \tag{17}$$

where $\alpha_i = \frac{1}{3}\left(1 + \frac{h_i}{h_{i-1}+h_i}\right)$ and thus Eq. 17 gives the weighted harmonic mean between Δ_{i-1} and Δ_i so that the relative spacing between the data points are considered. Eq. 17 is only valid (for preserving monotonicity) if $\Delta_{i-1}\Delta_i > 0$; that is, if Δ_{i-1} and Δ_i have the same sign and are distinct from zero. If this is not the case one sets d_i to zero. This should never be the case in our context, however.

The end points can be handled in different ways. The simplest solution, which we use in this study, is to use the one-sided finite differences:

$$d_1 = \Delta_1, \mathrm{d}_n = \Delta_{n-1} \tag{18}$$