# Science maps for information retrieval

Bascur Cifuentes, J.P.

# Science maps for information retrieval

by

Juan Pablo Bascur Cifuentes

Universiteit
Leiden
The Netherlands

Leiden, 2026

# Science maps for information retrieval

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr. S. de Rijcke,
volgens besluit van het college voor promoties
te verdedigen op woensdag 21 januari 2026
klokke 16:00 uur
door
Juan Pablo Bascur Cifuentes
geboren te Providencia, Chile
in 1987

**Promotores**
Prof.dr. L.R. Waltman
Prof.dr. S. Verberne

**Co-promotor:**
Dr. N.J.P. van Eck

**Promotiecommissie:**
Prof.dr. B.A. Barendregt (Voorzitter/Decaan Graduate School)
Prof.dr. R.J.W. Tijssen
Prof.dr. C.K. Kreutz (Technische Hochschule Mittelhessen)
Prof.dr. G. Cabanac (Université de Toulouse)
Dr. T. Velden (Deutsches Zentrum für Hochschul- und Wissenschaftsforschung)

# Contents

# List of Figures

# List of Tables

# Acknowledgments

Doing a PhD has been a life-changing experience. During these eight years, I have struggled, grown, and changed in ways I could never have imagined. But I could not have done this alone. Throughout this time, I have been supported, trusted, and taught by people who helped me learn things I could never have learned by myself.

I have so much to be grateful for, so I decided to organize this in a more manageable and structured way:

- **Ludo Waltman** and **Nees Jan van Eck**, who accepted me as a PhD student and had faith in my abilities.
- **Suzan Verberne**, who joined as my supervisor, believed in me, and provided critical support throughout my research.
- **Leiden University**, for making my burnout recovery easier.
- My partners **Roel van der Ploeg** and **Pepijn Stoop**, for their constant support during this journey.
- The **State of the Netherlands**, for giving me the opportunity to emigrate and integrate into a new society where I feel much more comfortable with the culture, the people, and the systems, and where I have found far greater opportunities for professional growth.
- **Suzan Verberne**, **Rodrigo Costas**, **Vincent Traag**, and **Ed Noijons**, for helping me find paid positions, which I desperately needed to sustain my self-funded PhD.
- For their rich academic discussions that shaped my research, I thank **Ludo Waltman**, **Nees Jan van Eck**, **Suzan Verberne**, **Rodrigo Costas**, **Vincent Traag**, **Ismael Rafols**, **Alfredo Yegros**, **Thed van Leeuwen**, **Martijn Visser**, **Jonathan Dudek**, **Wout Lamers**, and **Qianqian Xie**.
- **Karin den Dulk** and **Petra van der Weel**, for their kind support with administrative matters.
- Finding a postdoctoral position is never easy, and I am grateful to **Maxime Holmberg Sainte-Marie** for his support during that search.

During these years, I had the privilege of meeting many people in the academic environment whose friendship and companionship gave meaning and joy to my work. I am grateful to:

- My fellow CWTS researchers **Sonia Mena**, **Jonathan Dudek**, and **Mark Neijssel**.
- My fellow CWTS and LIACS PhD colleagues **Zhichao Fang**, **Ana Parrón**, **Wout Lamers**, **Alex Brandsen**, **André Brasil**, **Arian Askari**, **David Lindevelt**, and especially **Qianqian Xie**.
- My fellow visiting scholars **Zhentao Liang**, **Fabiano Borges**, **Gianmarco Spinaci**, **Thamyres Choji**, **Honami Numajiri**, **Jingwen Zhang**, and especially **Alause Pires**.
- My colleagues from the wider academic community **Wenjing Chu**, **Jing Wang**, **Daniëlle van Rijk**, **Felipe Castillo**, **Julia Eckert**, **Hiu Ching Hung**, **Eveline de Boer**, **Qinyu Chen**, **Satya Almasian**, and **Ylenia Casali**.

There is a world of people I am not mentioning simply because I cannot recall everyone right now. But please know that even if I have not named you, your actions have had a huge impact on my life. **Thank you.**

# Quote

*We are drowning in information but starved for knowledge*
— John Naisbitt, Megatrends

# Summary in English

Science maps are a widely used tool in scientometric analysis. One of their main advantages is that they reveal the structure of data, which for an analyst can both reveal unknown academic topics and address their own blind spots. These strengths make them well suited for information retrieval tasks, and researchers frequently employ science maps for this purpose. However, most existing research on science maps focuses on the accuracy of topic detection, while much less attention has been paid to their capabilities for information retrieval. This dissertation addresses this gap by exploring one overarching question: What is the effectiveness of science maps for information retrieval, and how can we enhance it?

The dissertation consists of an introduction, four Chapters (Chapters 2–5) that answer subquestions of the overarching question, and a conclusion. Each of these four subquestion Chapters is based on a peer-reviewed publication.

**Chapter 1: Introduction**

This Chapter defines what science maps are and to which information retrieval tasks they relate. It situates the research within the broader academic literature, introduces the research questions addressed in later Chapters, and presents the additional contributions of the dissertation beyond the research questions.

**Chapter 2: An interactive visual tool for scientific literature search: Proposal and algorithmic specification**

This Chapter addresses the question "How can science maps be designed to support information retrieval?" by proposing a tool that integrates science maps with an interactive retrieval process. The tool enables users to iteratively "scatter" a set of documents into clusters of related and then "gather" selected clusters into a refined subset. In addition, we developed an algorithm to position clusters in a visualization by minimizing the empty space while preserving the meaning of the distances between the clusters.

**Chapter 3: Academic information retrieval using citation clusters: In-depth evaluation based on systematic reviews**

This Chapter addresses the question "How effective are science maps for making systematic reviews?". In the evaluation we modeled information retrieval as an iterative process of selecting clusters and generating subclusters, using different user models with varying preferences for recall and precision. The results showed that science maps outperformed the boolean queries for about half of the reviews. This indicates that science maps are best used as a complement to, rather than a replacement for, other retrieval tools.

**Chapter 4: Which topics are best represented by science maps? An analysis of clustering effectiveness for citation and text similarity networks**

This Chapter addresses the question "Do science maps represent some topics better than others?" using Medical Subject Headings as the ground truth for topics. We measured clustering effectiveness for each topic individually and then aggregated by topic category. The best-represented categories were Organisms and Diseases, while Geographical entities were the least well represented. A topic was considered well clustered if its documents were concentrated in a few clusters and those clusters contained few unrelated documents. The evaluation was done on both citation-based and text-similarity-based maps across three different granularities. The analysis also showed that each category had similar clustering effectiveness on both types of maps.

**Chapter 5: Use of diverse data sources to control which topics emerge in a science map**

This Chapter addresses the question “How can the representation of specific topics be improved in a science map?” by comparing the performance of topics across science maps created with different data sources. We compared eight sources: text similarity, citations, co-authorship, patents, policy documents, and several forms of social media. The evaluation method is similar to Chapter 4, but modified to facilitate the comparison of a higher number of maps. A key modification was to compare the quality of maps for a given topic only at granularities where the topic was represented by the same number of clusters in both maps, which made the comparison more straightforward and flexible. Results showed that different sources could shift which topic categories performed best, but overall clustering quality tended to decrease when moving beyond text and citation networks. However, this performance loss could be mitigated by merging data sources.

**Chapter 6: Conclusion**

The Chapter summarizes the findings of the research questions and answers the overarching research question. It emphasizes that science maps can effectively support information retrieval, particularly when combined with other tools, and that science maps vary in their suitability across academic topics. The Chapter also outlines directions for future work, including how to obtain better performance from diverse data sources, the use of large language models, and the importance of prototyping and software sustainability.

# Summary in Dutch

*Science maps* zijn een veelgebruikt hulpmiddel in scientometrische analyses. Een van hun belangrijkste voordelen is dat ze de structuur van gegevens zichtbaar maken, wat voor een analist zowel onbekende academische onderwerpen kan onthullen als eigen blinde vlekken kan blootleggen. Deze sterke punten maken ze goed geschikt voor Information Retrieval (IR), en onderzoekers maken hier dan ook vaak gebruik van. Toch richt het meeste bestaande onderzoek naar science maps zich op de nauwkeurigheid van onderwerpdetectie, terwijl veel minder aandacht is besteed aan hun mogelijkheden voor IR. Dit proefschrift behandelt deze leemte door één overkoepelende vraag te onderzoeken: Wat is de effectiviteit van science maps voor IR, en hoe kunnen we deze verbeteren?

Het proefschrift bestaat uit een inleiding, vier hoofdstukken (Hoofdstukken 2–5) die deelvragen van de overkoepelende vraag beantwoorden, en een conclusie. Elk van deze vier hoofdstukken is gebaseerd op een peer-reviewed publicatie.

**Hoofdstuk 1: Inleiding**

Dit hoofdstuk definieert wat science maps zijn en met welke IR-taken ze verband houden. Het plaatst het onderzoek binnen de bredere academische literatuur, introduceert de onderzoeksvragen die in latere hoofdstukken worden behandeld, en presenteert de aanvullende bijdragen van het proefschrift naast de onderzoeksvragen.

**Hoofdstuk 2: Een interactieve visuele tool voor wetenschappelijke literatuurzoektocht: voorstel en algoritmische specificatie**

Dit hoofdstuk behandelt de vraag "Hoe kunnen science maps worden ontworpen ter ondersteuning van IR?" door een tool voor te stellen die science maps integreert met een interactief retrievalproces. De tool stelt gebruikers in staat om iteratief een set documenten te verspreiden over clusters en vervolgens geselecteerde clusters te verzamelen tot een verfijnde subset (het 'scatter-gather' paradigma). Daarnaast hebben we een algoritme ontwikkeld om clusters te visualiseren door lege ruimte te minimaliseren en tegelijkertijd de betekenisvolle afstanden tussen de clusters te behouden.

**Hoofdstuk 3: Academische IR met behulp van citatieclusters: diepgaande evaluatie op basis van systematische reviews**

Dit hoofdstuk behandelt de vraag "Hoe effectief zijn science maps bij het uitvoeren van systematische reviews?". In de evaluatie hebben we IR gemodelleerd als een iteratief proces van het selecteren van clusters en het genereren van subclusters, met verschillende gebruikersmodellen die uiteenlopende voorkeuren hadden voor recall en precisie. De resultaten toonden aan dat science maps beter presteerden dan de booleaanse zoekopdrachten bij ongeveer de helft van de reviews. Dit wijst erop dat science maps het best gebruikt kunnen worden als aanvulling op, in plaats van als vervanging van, andere zoekhulpmiddelen.

**Hoofdstuk 4: Welke onderwerpen worden het best weergegeven door science maps? Een analyse van clustereffectiviteit voor citatie- en tekstsimilariteitsnetwerken**

Dit hoofdstuk behandelt de vraag "Vertegenwoordigen science maps sommige onderwerpen beter dan andere?" met behulp van Medical Subject Headings als onderliggend systeem voor onderwerpen. We hebben de effectiviteit van clusters per onderwerp afzonderlijk gemeten en vervolgens per themacategorie geaggregeerd. De best geclusterde categorieën waren Organismen en Ziekten, terwijl Geografische entiteiten het minst goed werden weergegeven. Een onderwerp werd als goed geclusterd beschouwd als de bijbehorende documenten geconcentreerd waren in enkele clusters en die clusters weinig niet-gerelateerde documenten bevatten. We hebben de evaluatie uitgevoerd op science maps die gebouwd zijn zowel op citaties en inhoudelijke overeenkomst tussen documenten, op drie verschillende granulariteitsniveaus. Uit onze analyse bleek dat elke categorie vergelijkbare clustereffectiviteit vertoonde op beide typen science maps.

**Hoofdstuk 5: Gebruik van diverse databronnen om te sturen welke onderwerpen verschijnen in een wetenschapskaart**

Dit hoofdstuk behandelt de vraag "Hoe kan de weergave van specifieke onderwerpen in een science map worden verbeterd?" door de prestaties van onderwerpen te vergelijken tussen science maps die zijn opgebouwd met verschillende databronnen. We hebben acht soorten bronnen werden vergeleken: inhoudelijke overeenkomst, citaties, co-auteurschap, octrooien, beleidsdocumenten en verschillende vormen van sociale media. De evaluatiemethode lijkt op die van hoofdstuk 4, maar is aangepast om de vergelijking van een groter aantal science maps mogelijk te maken. Een belangrijke aanpassing was dat de prestaties van een onderwerp alleen werden vergeleken bij granulariteiten waarbij de verschillende kaarten hetzelfde aantal onderwerpclusters opleverden. De resultaten toonden aan dat verschillende bronnen konden verschuiven welke themacategorieën het best presteerden, maar dat de algehele clusterkwaliteit vaak afnam bij het gebruik van andere bronnen dan tekst- en citatienetwerken. Dit kwaliteitsverlies kon echter worden beperkt door databronnen te combineren.

**Hoofdstuk 6: Conclusie**

Dit hoofdstuk vat de bevindingen van de onderzoeksvragen samen en beantwoordt de overkoepelende onderzoeksvraag. Het benadrukt dat science maps effectief kunnen bijdragen aan IR, vooral in combinatie met andere tools, en dat hun geschiktheid verschilt per academisch onderwerp. Het hoofdstuk schetst ook richtingen voor toekomstig onderzoek, waaronder hoe betere prestaties kunnen worden bereikt met diverse databronnen, het gebruik van grote taalmodellen, en het belang van prototyping en software sustainability.

# About the author

Juan Pablo Bascur Cifuentes (1987) was born in Providencia, Chile. He graduated from Colegio Cumbres high school in 2005 and went on to study Biochemistry at the Andrés Bello National University in Santiago between 2008 and 2016, where he obtained his BSc and MSc degree cum laude. During this period, he also worked as a private tutor for high school and bachelor students. His master's dissertation, supervised by Daniel Eduardo Almonacid, focused on the identification of enzyme superfamilies significantly mutated in cancer and the characterization of their analogous mutations. Between 2013 and 2016, he also worked as a researcher on a Fondecyt project, where he developed network-based methods to predict the biological function of genes. After completing his studies, he worked as a machine learning researcher at the company WriteWise (2017–2018), where he developed and implemented functionalities for academic writing assistance software.

In 2018, he moved to Leiden in the Netherlands to pursue a PhD in Scientometrics at Leiden University, which he will defend in January 2026. He was supervised by Ludo Waltman and Nees Jan van Eck from the Centre for Science and Technology Studies and Suzan Verberne from the Leiden Institute of Advanced Computer Science. During his doctorate, his publications and conference presentations have focused on the science of science, information retrieval, machine learning, and computational social science.

Alongside his doctoral research, he has been actively engaged in teaching, research, and academic service. Since 2021 he has worked as a teaching assistant at the Leiden Institute of Advanced Computer Science, contributing to courses in information retrieval, text mining, data science, and artificial intelligence. He has also served as a social media analyst for two research projects with Delft University of Technology and Fondazione Bruno Kessler, as a scientometrics analyst for three CWTS B.V. projects, and as a Python workshop lecturer. In addition, he has been an editor for the Leiden Madtrics blog, an organizer of a workshop at the International School and Conference on Network Science, and provided extensive peer review for multiple journals.

# Chapter 1

# Introduction

Navigating academic literature has never been easy, and it is becoming harder each year due to the accelerating rate of literature production [6, 26, 115]. Traditional search methods, such as keyword-based queries, work well for users familiar with the topic, but leave them unaware of blind spots [106, 169]. On the other hand, science maps are tools to visualize the relationships between academic documents, which function as an overview of the research landscape. Science maps can support traditional search by showing overlooked connections and allowing expert knowledge to improve the search. Despite this, science maps are not part of information seeking manuals or studies [32, 95]. Given this situation, it would be beneficial to get a better understanding of the performance of science maps for information retrieval tasks. In this dissertation, we do this by providing evidence for the benefits and drawbacks of using science maps for specific information retrieval tasks, and by exploring ways for using them more effectively.

## 1.1 Use of science maps

Science maps are tools to visually explore the relations between objects of interest in a large collection of documents [25, 38, 127]. Their most typical use is in studying the structure of a scientific field. For example, each year there are countless bibliometrics studies of different academic fields that have science maps as their core method of study. Among other things, these studies delimit the field [185], identify topical trends, research groups, and key actors and events. The target audience of these studies are typically members of the same field, as it allows them to know where their research stands in relation to their colleagues. Beyond studying the structure of a scientific field, policymakers and research administrators also use science maps to characterize national research output, evaluate the impact of funding programs, and identify competitive advantages and weaknesses in different scientific areas.

Science maps belong to a broader family of tools that attempt to represent and summarize large collections of data in a structured and interpretable manner. These tools are used both for analysis and communication. For example, Latent Dirichlet Allocation [24], a method to algorithmically generate topic models, is used to identify topics in documents according to the co-occurence of words in documents. Similarly, conceptual maps [58], which are manually constructed diagrams of relations between concepts, are used in education to explain how concepts relate to each other.

Science map visualizations are typically a bubble chart that represents a network, where the bubbles (i.e. nodes) represent objects of interest. These nodes are typically either academic researchers, journals, organizations, countries, terms from the documents, or clusters of documents with a topical label. Figure 1.1 is an example of a science map that visualizes clusters of documents, and Figure 1.2 is an example of a science map that visualizes authors.

The bubbles (nodes) in a science map are placed in a two dimensional plane, sometimes connected by the network edges, where the spatial position and edges represent the relation of the nodes with

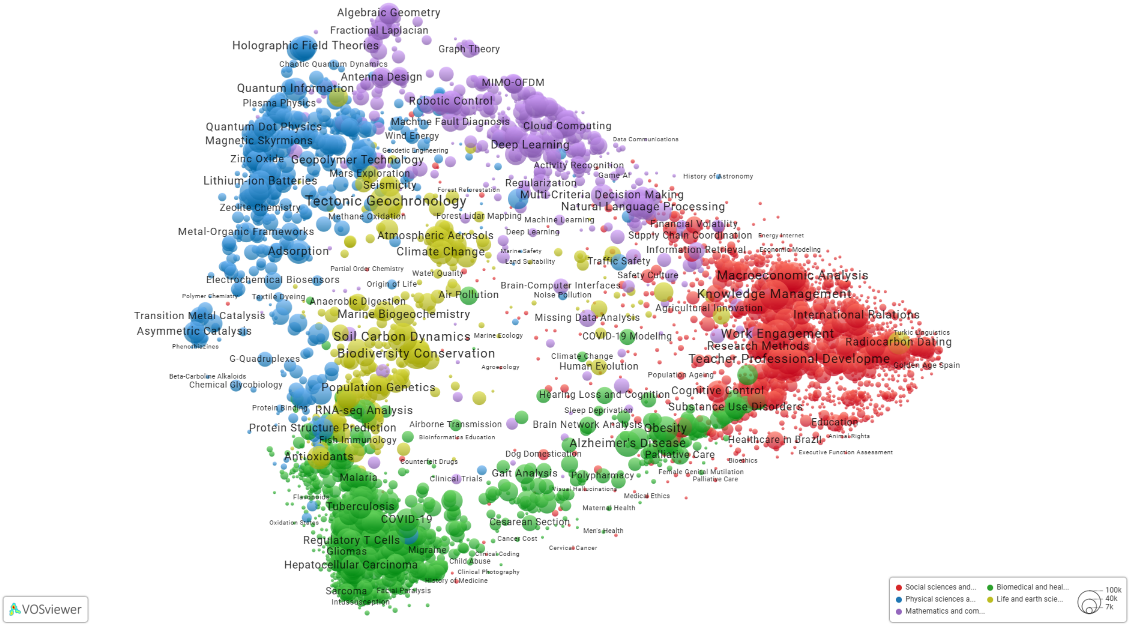


Figure 1.1: Example of a science map that visualizes clusters of documents. Spatial proximity indicates the intensity of intercluster citations, and color indicates the field of science. Source [159]

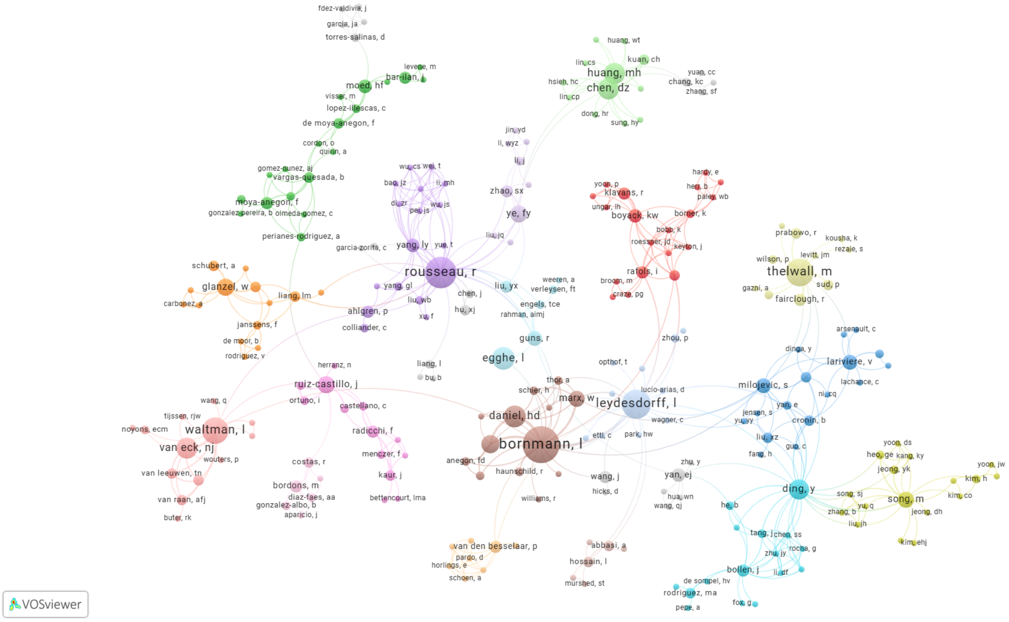


Figure 1.2: Example of a science map that visualizes authors. Links and spatial proximity indicate co-authorships, and color indicates that they belong to the same cluster in the network. Source [155]

each other. Other properties, such as the color and size, are used to represent other properties of the node. This representation allows users of science maps to quickly find connections between the nodes. For example, authors that are placed together (Figure 1.2) probably belong to the same research group or work on the same topic.

## 1.2 Document clusters in science maps

There are many types of science maps, but in this dissertation we only focus on the ones that use clusters of documents as nodes in their visualizations [38, 151, 164]. We do this because the use of these maps in information retrieval is straight-foward, as a user can just retrieve the documents that belong to a given cluster. Typically, the clusters are created the following way: First, the collection of documents to be visualized by the science map is turned into a network, where the nodes are the documents and the links are based on the metadata of the documents (see below). Then, a clustering algorithm is run to find communities of documents in the network. A frequently used algorithm is the Leiden algorithm, which requires a resolution value to be provided to determine the granularity of the clusters [153]. These two steps are the ones we study the most in this dissertation. Then, the next step before visualization is to assign labels to the clusters, which typically is done with conventional text processing techniques (like word count vectors [111]) rather than advanced (like text embeddings [50]) or field-specific ones (like named entity recognition of diseases in health records [168]). For example, Waltman and van Eck [164] did the former, while van Eck and Waltman [159] did the latter. Finally, the clusters are visualized as a network of clusters using the approach described in Section 1.1.

There are three types of networks that are typically used for document clustering. The first and most common type is the citation network. In its most simple implementation, called direct citations network, the edges of the network are the citation links between documents. Using citations has the advantage that a citation is an explicit intellectual link made by academics. There are three other implementations of citation networks:

- Extended direct citation network: It includes citations to documents that are not part of the science map [165].
- Co-citation network: An edge in the network indicates that the two linked documents are cited by the same third document [145].
- Bibliographic coupling network: An edge in the network indicates that the two linked documents cite the same third document [96].

The edges of co-citation and bibliographic coupling networks can have different weights according to how many documents they are being cited by or are citing together.

The second type of network used for document clustering is the similarity network, where similarity typically is text similarity. The weight of the edge between two documents is the strength of the text similarity, and the text used to determine the text similarity typically consists of the titles and abstracts of the documents because this data tends to be readily available. The methods used to calculate the similarity are diverse, but it is notable that advanced methods, like text embedding, tend to not be used, likely due to bibliometricians preferring explainability over performance. The text similarity network has an advantage over the citation network in that the similarity between any pair of documents can be calculated, which helps when citation links between the documents are sparse, for example citation networks where the documents are very new or very few. There are also hybrid networks where both citations and text similarity contribute to the weights of the edges [4, 27].

The third type of network is the co-occurrence network. In this network, an edge between two documents indicates that the documents share something, like an author, or are placed together in some context, like being cited by a patent. It is worth mentioning that citation and text networks can also be seen as co-occurrence networks (like in the previously mentioned co-citation and bibliographic coupling, or by sharing a word in the text). Beyond these, some of the most common elements used to connect documents in a co-occurrence network are institutions, policy documents, social media posts and key words [45, 101]. Co-occurrence networks are used to answer specific questions related to the entities that connect the documents, for instance how social media connects academic documents.

## 1.3 Information retrieval with clusters

The field of information retrieval studies how to find relevant information. Over time, research in this field has shifted away from using clustering to find information due to the emergence of well performing ranking algorithms. The pivotal point came with the emergence of PageRank [29], from Google, which was able to rank web pages not only according to the relevance of their textual content to a query, but also based on the number and weight of the hyperlinks pointing at it. Even so, cluster-based information retrieval should not be ignored because there are tasks that are better served by clustering than by a ranked list of results [176]. One of the most relevant uses is diversification of search results [72]: In the example of PageRank, it is possible for a query to match several topics. If only one of the topics appears at the top of the result list, then there is the risk that this is not the topic that the user is looking for. For example, if the user is looking for the homepage of the company Jaguar, but all the results are about the animal. With results diversification, the results can be clustered by topic and then ensure that all the topics are represented at the top of the list. The search engine Carrot2 goes one step further [65] and makes the results topic clusters available to the user. Another use of clusters is query expansion [108], where the user does not know all the relevant query terms for the search and then the system suggests new terms based on the ones that the user already provided.

As a general rule, clustering in information retrieval is most useful at assisting in complex tasks that require several steps, instead of simple tasks like finding a known individual document [71]. These complex tasks require the user to explore the results of each step so as to decide on the next step. Clustering also supports complex tasks by giving tools to the users to expand or reduce their search results in a sensible way. An excellent example of clustering for complex multi-step tasks is the Scatter-Gather method [48], which was originally proposed to facilitate the navigation of news articles. It creates clusters based on the text similarity between the documents, then labels the clusters, and then lets the user select which clusters might contain documents with their topic of interest. Then the method creates a new set of clusters using only the documents in the previously selected clusters, and then the process repeats until the user identifies a cluster labeled with their topic of interest. This approach is ideal when the user struggles to articulate effectively their topic of interest, when they do not know how to find their topic of interest in the documents, or when the documents categorization is lacking.

In this dissertation, we hypothesize that the Scatter-Gather method is ideal for academic search for two reasons:

- Because academic users tend to have complex information needs [136].
- Because academic documents categorization tends to be lacking due to the rate of emergence of new topics and research questions, and the difficulty to maintain a classification system [3, 180].

## 1.4 Bibliometrics enhanced information retrieval

Academics have used bibliometrics to enhance information retrieval through the use of citations [30, 63, 114]. The closest to the Scatter-Gather method is the tool CitNetExplorer [157]. This tool creates a network of citations between documents, and it can also identify clusters within the network. Its user interface facilitates selection of documents based on the clusters they belong to. It can also create a new clustering solution based on selected documents, including documents selected using the clusters, which allows the users to easily follow the Scatter-Gather method. It also facilitates selecting additional documents based on their citation links to the currently selected documents, something that is arguably a helpful addition because the Scatter-Gather method only allows to remove documents. However, instead of cluster-based methods, academics tend to use citations for information retrieval using a method called citation snowballing [23]. Snowballing consists of selecting one or more documents, gathering the documents that cite them or are cited by them, and then repeating the process with this new set of documents until the user is satisfied. To limit the number of selected documents, there is usually a limit on how many expansion cycles to make or a minimum threshold on the number of selected documents a new document must be connected to. For example, Janssens and Gwinn [89] found that, for finding the relevant documents of systematic reviews, it was most convenient to add documents that are co-cited by multiple seed documents and to use the number of co-citations as a threshold for relevance.

Additionally, there are tools that visualize the structure of search results, either by text similarity (like Open knowledge maps [122] and Iris.ai [88]) or citation (like Inciteful [87], Litmaps [107], Connected papers [44] and Research rabbit [99]). A difference between these tools and science maps is that these tools operate at a very small scale, visualizing individual documents. This has the advantage that the tools work in a way that is intuitive and easy-to-understand for users, but it misses the big picture view that is provided by working with big clusters of documents. For example, most of the clusters in Figure 1.1 have between 1,000 and 100,000 documents. CitNetExplorer found a middle point between visualizing clusters and visualizing individual documents by visualizing only the most important documents, which are identified by the properties of their nodes in the citation network. The clusters are kept in the back end of the system and the visualization indicates the clusters of the visualized documents. Another difference between science maps and the above-mentioned tools is that the latter tend to create their network of documents starting from seed documents provided by the users, which does not allow the users to know what they are missing. This is particularly problematic where there are communities of relevant documents disconnected from each other, either by lack of citations or lack of similar language [2, 80, 134]. For example, the collection of documents about academic information retrieval has a low citation connectivity between its biomedicine and computer science communities, and a user might never be aware of this if they use the above-mentioned tools with seed documents from one community only.

In summary, information retrieval based on science maps uses similar concepts and methods to bibliometric enhanced information retrieval, but it also has potential advantages for information retrieval that differentiates it from the latter. Given this, we identify the lack of knowledge on the performance of science maps for information retrieval as a research gap that we will attempt to fill in this dissertation.

## 1.5 Research questions

Our motivation for the research presented in this dissertation is to explore the potential of science maps to enhance academic search by saving time, improving knowledge discovery, and providing a more complete picture of the state of the art in the literature. The benefits of science maps may be especially significant for early career researchers and citizen scientists, who tend to be less familiar with their research field. Our research also adds value to the field of science mapping because it provides a new application area for the knowledge generated by the field. The way we see it, a research agenda for this purpose should evaluate the effectiveness of science maps for information retrieval tasks and also improve this effectiveness. Therefore, in this dissertation, our overarching research question is: **What is the effectiveness of science maps for information retrieval, and how can we enhance it?** We address this overarching research question by answering a number of more specific research questions, and each of them is the topic of a separate chapter in this thesis:

- Research question 1 (Chapter 2): The first step in our research is to understand how to use science maps for information retrieval. Our research question is: **How can science maps be designed to support information retrieval?** Here, we propose a system, named SciMacro

(Science Macroscope), for interacting with science maps that serves information retrieval tasks based on the Scatter-Gather method. We find no significant hindrances for the implementation. Figure 1.3 shows a screenshot of the graphical user interface that we created for SciMacro, whose source code is publicly available [15].

- Research question 2 (Chapter 3): To evaluate science maps for information retrieval tasks, we start by addressing the research question: **How effective are science maps for making systematic reviews?** Here, we evaluate the performance of science maps at retrieving the relevant documents of systematic reviews, using the Boolean queries of systematic reviews as baseline. We find that science maps are able to outperform the baseline for about half of the systematic reviews.
- Research question 3 (Chapter 4): We also consider whether the performance of a science map depends on the academic topic of the task. Our research question is: **Do science maps represent some topics better than others?** Here, we evaluate the performance of science maps at creating clusters for topics, using Medical Subject Headings as topics. We find that both text and citation based maps cluster the topics that belong to ontological categories of topics "Organisms" or "Diseases" much better than the other topics.
- Research question 4 (Chapter 5): Finally, we consider whether we can manipulate a science map to perform better at a given academic topic. Our research question is: **How can the representation of specific topics be improved in a science map?** Here, we evaluate if the use of different kinds of document networks influences which topics are clustered better than others. We find that such an influence indeed exists, but also that the topics from most ontological categories of topics decreased their performance in the new networks relative to text or citation networks and that performance can be improved by merging different network types.

Research questions 1 and 4 investigate the use and improvement of science maps for information retrieval, while research questions 2 and 3 evaluate their effectiveness. The progression is as follows: First, we conceptualize how science maps can support information retrieval (RQ1). Based on this, we then evaluate their performance (RQ2) and find it to be uneven. To understand this variation, we examine whether topic differences play a role (RQ3) and confirm that they do. Finally, we investigate whether performance for underrepresented topics can be improved by changing the data source of the map (RQ4). We conclude this dissertation in Chapter 6, where we summarize key findings and discuss future research directions for science maps for information retrieval. It is worth noting that this dissertation does not include experiments with real users. Although such experiments could help answer the research questions, they would require resources and expertise beyond the scope of this dissertation.

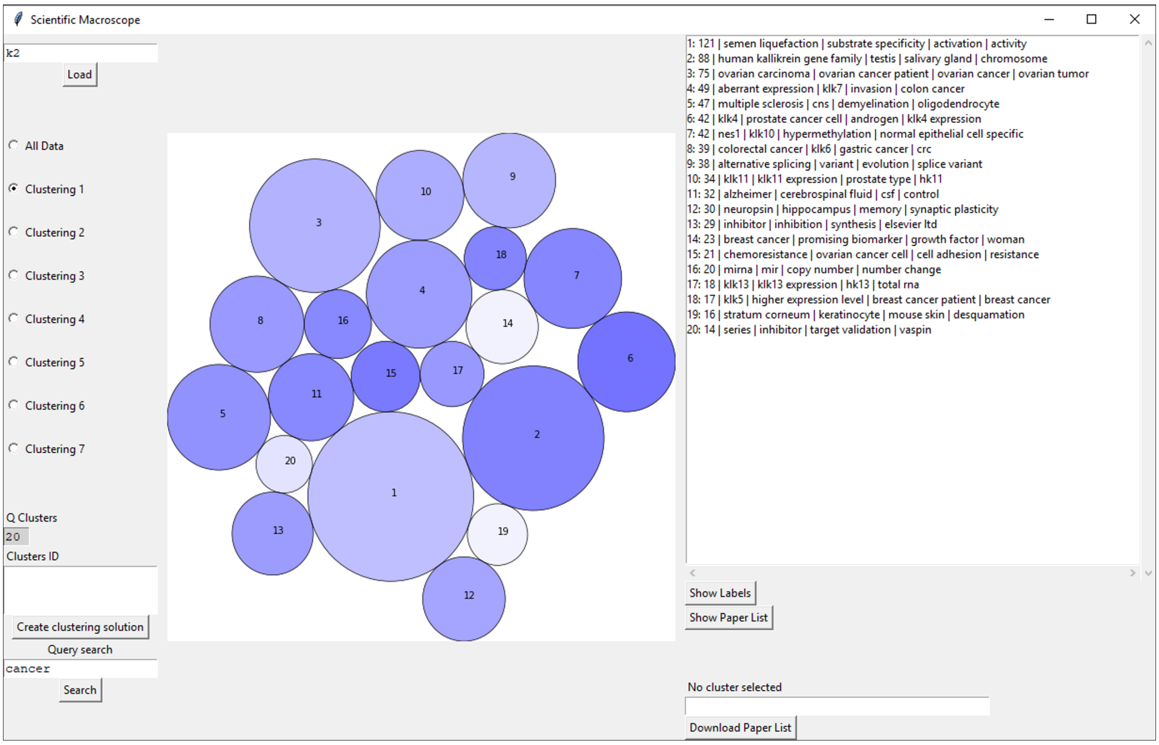
## 1.6 Main contributions

While this dissertation focuses on answering the research questions, we also made other additional contributions to the field. We divide the contributions between resource contributions and methods contributions.

### 1.6.1 Resource contributions

These are the resources that we generated during our dissertation, and they facilitate either the design of science mapping tools or the execution of information retrieval experiments:

- In Chapter 2, we introduce a science mapping tool prototype that follows the Scatter-Gather principles, including an algorithm that places the bubbles close to each other while also preventing overlapping and minimizing empty space. We have made the code publicly available [15].



k2

○ All Data
● Clustering 1
○ Clustering 2
○ Clustering 3
○ Clustering 4
○ Clustering 5
○ Clustering 6
○ Clustering 7

Q Clusters: 20

Clusters ID:

Query search: cancer

```
1: 121 | semen liquefaction | substrate specificity | activation | activity
2: 88 | human kallikrein gene family | testis | salivary gland | chromosome
3: 75 | ovarian carcinoma | ovarian cancer patient | ovarian cancer | ovarian tumor
4: 49 | aberrant expression | klk7 | invasion | colon cancer
5: 47 | multiple sclerosis | cns | demyelination | oligodendrocyte
6: 42 | klk4 | prostate cancer cell | androgen | klk4 expression
7: 42 | nes1 | klk10 | hypermethylation | normal epithelial cell specific
8: 39 | colorectal cancer | klk6 | gastric cancer | crc
9: 38 | alternative splicing | variant | evolution | splice variant
10: 34 | klk11 | klk11 expression | prostate type | hk11
11: 32 | alzheimer | cerebrospinal fluid | csf | control
12: 30 | neuropsin | hippocampus | memory | synaptic plasticity
13: 29 | inhibitor | inhibition | synthesis | elsevier ltd
14: 23 | breast cancer | promising biomarker | growth factor | woman
15: 21 | chemoresistance | ovarian cancer cell | cell adhesion | resistance
16: 20 | mirna | mir | copy number | number change
17: 18 | klk13 | klk13 expression | hk13 | total rna
18: 17 | klk5 | higher expression level | breast cancer patient | breast cancer
19: 16 | stratum corneum | keratinocyte | mouse skin | desquamation
20: 14 | series | inhibitor | target validation | vaspin
```

No cluster selected

Figure 1.3: Screenshot of the graphical user interface of SciMacro. On the left it allows the user to load the documents they want to cluster, indicate how many clusters they want, select clusters for the next clustering step, go back and forth between the clustering steps and search for words in the clusters whose frequency is represented by the color of the clusters. On the right it provides a description of the clusters, shows the documents inside the clusters, and allows downloading the documents in the selected cluster. Source code of the interface [15]

- In Chapter 3, we have developed [14] a manually curated version of the search queries dataset from Scells et al. [139] that is compatible with the PubMed API query grammar and that can be used for information retrieval experiments.
- In Chapters 3, 4 and 5, the source code of all the experiments is provided, along with a modified version of the experimental data that allows replication of the experiments. The data is modified to prevent legal violations.

### 1.6.2 Methodological contributions

In the current dissertation we evaluated the performance of science maps for information retrieval and explored ways of improving it.

- We proposed a user model that assumes a user that has perfect knowledge about the location of relevant documents, and use it in Chapter 3, 4 and 5 to design original evaluation methods.
- We proposed two approaches to select which are the relevant documents that should be used to evaluate science maps for different tasks: In Chapter 3, we proposed to use the included and excluded documents in a systematic review. In Chapter 4 and 5, we proposed to use the documents labeled with MeSH terms to evaluate academic topics and allowed the documents to inherit MeSH terms higher up in the ontology.
- We proposed how to evaluate the clusters in science maps in a way that manages the size disparity between clusters and number of relevant documents. In Chapter 3, we proposed to allow the user model to manage the granularity of the clusters. In Chapter 4 and 5, we proposed a method that aggregates the evaluations of multiple granularities.

## 1.7 List of publications

Chapters 2 to 5 in this thesis are based on the following publications.

**Chapter 2:** Juan Pablo Bascur, Nees Jan van Eck and Ludo Waltman. 2019. An interactive visual tool for scientific literature search: Proposal and algorithmic specification. Proceedings of the 8th International Workshop on Bibliometric-Enhanced Information Retrieval (BIR) Co-Located with the 41st European Conference on Information Retrieval (ECIR 2019), 76–87. https://ceur-ws.org/Vol-2345/paper7.pdf [16]

**CRediT author statement:**

- Juan Pablo Bascur: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.
- Nees Jan van Eck: Conceptualization, Methodology, Supervision, Writing – review & editing.
- Ludo Waltman: Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

**Chapter 3:** Juan Pablo Bascur, Suzan Verberne, Nees Jan van Eck and Ludo Waltman. 2023. Academic information retrieval using citation clusters: In-depth evaluation based on systematic reviews. Scientometrics, 128, 2895–2921. https://doi.org/10.1007/s11192-023-04681-x [17]

**CRediT author statement:**

- Juan Pablo Bascur: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.
- Suzan Verberne: Conceptualization, Methodology, Supervision, Writing – review & editing.
- Nees Jan van Eck: Conceptualization, Methodology, Software, Supervision, Writing – review & editing.
- Ludo Waltman: Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

**Chapter 4:** Juan Pablo Bascur, Suzan Verberne, Nees Jan van Eck and Ludo Waltman. 2025. Which topics are best represented by science maps? An analysis of clustering effectiveness for citation and text similarity networks. Scientometrics 130, 1181–1199. https://doi.org/10.1007/s11192-024-05218-6 [19]

**CRediT author statement:**

- Juan Pablo Bascur: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.
- Suzan Verberne: Conceptualization, Methodology, Supervision, Writing – review & editing.
- Nees Jan van Eck: Conceptualization, Methodology, Supervision, Writing – review & editing.
- Ludo Waltman: Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

**Chapter 5:** Juan Pablo Bascur, Rodrigo Costas, Suzan Verberne. 2024. Use of diverse data sources to control which topics emerge in a science map. arXiv. https://doi.org/10.48550/arXiv.2412.07550 [18]

**CRediT author statement:**

- Juan Pablo Bascur: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.
- Rodrigo Costas: Conceptualization, Writing – review & editing.
- Suzan Verberne: Conceptualization, Methodology, Supervision, Writing – review & editing.

# Chapter 2

# An interactive visual tool for scientific literature search: Proposal and algorithmic specification

## Abstract[1]

Literature search is a critical step in scientific research. Most of the current literature search tools present the search results as a list of documents. These tools fail to show the structure of the search results. To address this issue, we propose an interactive visual tool for searching scientific literature. This tool creates, labels and visualizes clusters of documents that may be of relevance to the user. In this way, it provides the user with an overview of the structure of the search results. This overview is intended to be understandable even to a user who has only a limited familiarity with the scientific domain of interest. We present the concept of our tool, show a case study of its use and describe the technical specifications of the tool. In particular, we provide a detailed specification of the algorithm that we use to visualize clusters of documents.

## 2.1 Introduction

Literature search is an essential part of any research project. Many of the current literature search tools (e.g. Google Scholar [66], Web of Science [41], Scopus [56] and Dimensions [51]) present the search results as a list of documents, without showing the structure of the results. Getting an understanding of the structure of the results, for instance by providing a breakdown of the search results into different research topics, can be useful for exploring the literature [1], especially for making serendipitous discoveries or for users that are new to a field of research.

There is some literature studying the idea of showing the structure of search results. An example is the recent work on a tool called PaperPoles [71], which uses citation links to create clusters of related papers. Various tools have also been made publicly available, some of them with a clear focus on literature search and others with a primary focus on bibliometric analysis. For instance, CiteSpace [39], CitNetExplorer [157] and Citation Gecko [163] can be used to visualize networks of

[1] This chapter is based on: Juan Pablo Bascur, Nees Jan van Eck and Ludo Waltman. 2019. An interactive visual tool for scientific literature search: Proposal and algorithmic specification. Proceedings of the 8th International Workshop on Bibliometric-Enhanced Information Retrieval (BIR) Co-Located with the 41st European Conference on Information Retrieval (ECIR 2019), 76–87. https://ceur-ws.org/Vol-2345/paper7.pdf [16]

citations between documents. Open Knowledge Maps [122] shows clusters of semantically-related papers. VOSviewer [156] presents visualizations of co-occurrence networks derived from papers (e.g. co-authorship links between authors, citation links between documents, or co-occurrence links between terms).

While these tools are helpful, some of them (e.g. CiteSpace, VOSviewer) were developed primarily for bibliometric analysis, not for literature search. Others (e.g. CitNetExplorer, Citation Gecko) have the limitation of showing search results only at the level of individual papers, not at aggregate levels. To overcome the limitations of currently available tools, we propose a new tool for literature search. This tool uses an interactive visual interface to show the structure of the search results. We make use of ideas and techniques that we also used in the development of other tools (i.e., VOSviewer and CitNetExplorer), but we now focus specifically on literature search rather than on bibliometric analysis. To some degree, the proposed tool resembles Open Knowledge Maps. However, by relying on the Scatter/Gather approach [48], the tool offers a higher level of interactivity, which facilitates the exploration of large document spaces.

This paper is divided into three parts: We first provide a description of the proposed tool (Section 2.2), we then present a case study demonstrating the use of the tool (Section 2.3) and finally we give a technical specification of the algorithms included in the tool (Section 2.4).

## 2.2 Description of the tool

Our proposed tool is based on the Scatter/Gather approach [48]. This approach consists of exploring a set of documents through multiple iterations of scattering and gathering. To scatter means creating clusters of documents and labeling them to understand their contents. To gather means selecting the clusters of interest, resulting in a new set of documents (Figure 2.1). The documents in our tool are scientific papers.

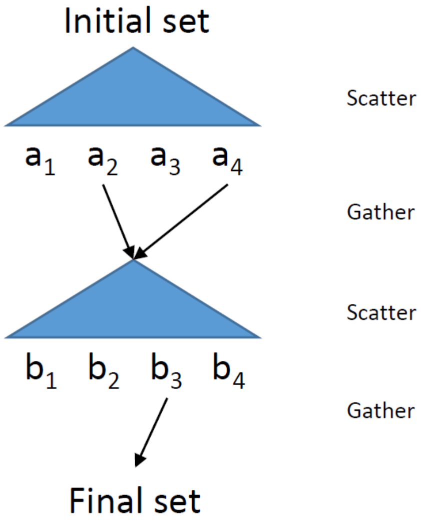


Figure 2.1: The Scatter/Gather approach. Figure inspired by Figure 1 of Cutting et al. [48]. The user scatters the initial set of documents into labeled clusters of documents (a1, a2, a3, and a4). Then she gathers the clusters she is interested in and creates a new set of documents. Then she scatters the new set into new clusters (b1, b2, b3, and b4). This process can continue a number of times.

Our tool scatters a set of papers into clusters. The clustering uses the citation links between papers. Each cluster is given a label. The label of a cluster consists of the ten noun phrases with the highest weighted frequency in the titles and abstracts of the papers in the cluster. The weighting considers the frequency of occurrence of the noun phrases in the focal cluster relative to other clusters. This clustering and labeling method is based on Waltman and Van Eck [164].

Our tool also visualizes the clusters to complement the labels. It visualizes the clusters as bubbles in a packed bubble chart. The size of the bubbles reflects the number of papers in the clusters and the distance between the bubbles approximately reflects the number of citation links between the clusters.

Our tool supports multiple iterations of scattering and gathering. The user can load the initial set of papers, choose the clusters to gather, choose the number of clusters to scatter, retrieve the papers in the clusters, and so on.

## 2.3 Case study of the tool

### 2.3.1 Set up

First, let us consider a user working with a traditional literature search engine for scientific literature, like Google Scholar. She has to come up with several search queries. She does not have a background in the academic field that she is looking into, so probably she will not come up with good queries. Also, she has no way to know if she is missing important papers or even entire subfields!

Second, let us assume instead that she uses a literature search engine that offers some very basic features for exploring the structure of the search results, like Web of Science. She can now see to which academic fields her search results belong. Despite of this, she still has basically the same problems as with Google Scholar.

Third, now let us assume that she uses our proposed tool for her literature search. For this example, we will follow her through all the steps of the search process. We will assume that she is interested in getting to know the scientific literature about the review process of grant proposals. For the initial set of papers, we will use the set of the cluster of scientometrics papers obtained using the algorithmic methodology employed at CWTS [164]. We believe that she would have used the same set because it covers her topic.

### 2.3.2 Example of the search process

The researcher retrieves the set of papers and chooses a value of 10 for the number of clusters in the first scattering. Then she sees the visualization (Figure 2.2A) and the labels (Table 2.1) of the clusters. From the labels, she sees that her topic of interest is in cluster 6. She also checks the labels of the clusters close to cluster 6 (clusters 0, 3, 5, 8 and 9). Their labels indicate that they do not relate to her topic of interest, so she only gathers cluster 6.

She chooses to have 5 clusters for the second scattering and sees the visualization (Figure 2.2B) and the labels (Table 2.2) of the clusters. Now the labels are more ambiguous, so she will have to also read the titles of the papers inside clusters to understand what the clusters are about. She suspects that her topic of interest is in clusters 1 and 2. From the visualization and the labels, she also sees that her topic could be in cluster 4. She reads the titles of the top 5 most cited papers in these three clusters (Tables 2.3, 2.4 and 2.5). She finally decides that she should start reading paper 3 from cluster 1 and papers 2 and 4 from cluster 2.

In this example, we have illustrated how our tool could improve scientific literature search. The key advantage of the proposed tool is that the user is informed about the way in which the scientific literature is organized. For instance, the user is able to see how a field is divided into subfields or topics. As a result of this, the user is able to discard papers unrelated to the topic of interest without the need to skim the titles of large numbers of individual papers. Instead, the user examines the labels of clusters and then decides to discard entire clusters that appear to be of no relevance. Also, the user does not need to try to come up with a detailed keyword query that identifies exactly the right papers. It is sufficient to be able to identify a broad set of papers that could potentially be of relevance. Within this broad set of papers, the papers of interest can then be found by drilling down into the right clusters.

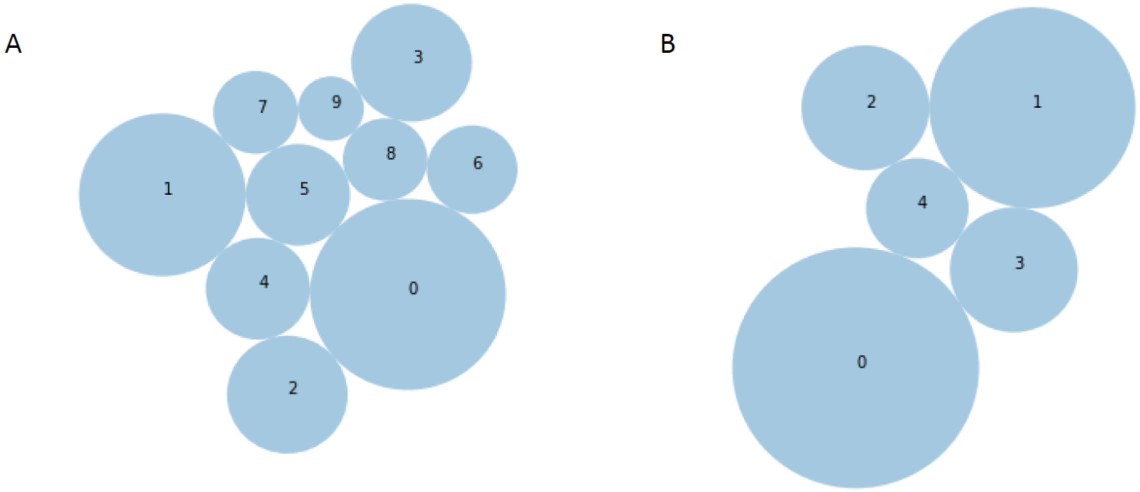


Figure 2.2: Visualization of clusters. The size of a cluster reflects the number of documents belonging to the cluster. Clusters that are strongly related (based on citation links) tend to be located close to each other. The numbers are the identifiers of the clusters. A: First scattering. B: Second scattering.

## 2.4 Technical specification

### 2.4.1 Clustering the documents

We cluster the papers by applying the Leiden algorithm to their citations links [153, 164]. The Leiden algorithm identifies clusters (or communities) of nodes within a network. We apply the Leiden algorithm to a directed network where the papers are the nodes and the edges are the citations between citing and cited papers. The Leiden algorithm has a resolution parameter that determines the number and size of clusters. To avoid requiring the user to set the resolution parameter manually, we developed a rule of the thumb that enables the user to specify the number of clusters $C$ that she wishes. According to this rule, the resolution parameter is chosen in such a way that the largest cluster includes between $N/(C-2)$ and $N/(C)$ papers, where $N$ is the total number of papers in the collection. To obtain the desired number of clusters after the clustering algorithm has been run, we keep the top $C$ largest clusters and merge them with the other smaller clusters. We merge the pairs of clusters that have the highest relatedness, which we define as $e(c_1, c_2)/(n(c_1) * n(c_2))$, where $c_1$ and $c_2$ are the clusters, $e(c_1, c_2)$ is the number of edges between two clusters and $n(c)$ is the number of papers in a cluster.

### 2.4.2 Labeling the clusters

We label clusters using the approach developed by Waltman and Van Eck [164]. This approach extracts cluster labels from noun phrases in the titles and abstracts of the papers belonging to a cluster. It labels a cluster using noun phrases that are common in the cluster and relatively uncommon in other clusters. The only modification that we make to the approach introduced in [164] is that we report 10 noun phrases instead of 5.

Table 2.1: Labels of the first scattering. Scattered from the cluster of scientometrics papers [164].

| ID | Top 10 noun phrases | Papers |
|---|---|---|
| 0 | hirsch \| h index \| g index \| citation distribution \| hirsch index \| index \| percentile \| variant \| google scholar \| calculation | 4344 |
| 1 | man \| gender difference \| scientific collaboration \| research collaboration \| woman \| co authorship network \| international committee \| gender \| medical journal editors \| icmje | 3154 |
| 2 | citation classic \| article type \| randomized controlled trial \| year survey \| gross domestic product \| study design \| pubmed database \| subspecialty \| population size \| medline database | 1652 |
| 3 | open access \| institutional repository \| open access publishing \| altmetric \| oa journal \| self archiving \| open access journal \| mendeley \| repository \| twitter | 1651 |
| 4 | author keyword \| nanotechnology \| patent citation \| patent \| chinese academy \| nanotechnology research \| nanoscience \| keywords plus \| productive journal \| uspto | 1231 |
| 5 | interdisciplinarity \| bibliographic coupling \| co word analysis \| research front \| aca \| map \| intellectual structure \| visualization \| co citation \| cluster | 1230 |
| 6 | peer review process \| rejection \| reviewer \| peer reviewer \| peer review \| review quality \| review process \| manuscript \| manuscript review \| peer review system | 932 |
| 7 | link analysis \| hyperlink \| web page \| inlink \| web link \| web site \| yahoo \| search engine \| web impact factor \| link count | 816 |
| 8 | marketing \| operations management \| management journal \| citation error \| finance journal \| rpys \| business school \| quotation error \| management discipline \| reference accuracy | 810 |
| 9 | economics department \| economist \| economics journal \| academic economist \| economic research \| economic \| jel \| american economic review \| economics profession \| top economics journal | 492 |

Table 2.2: Labels of the second scattering. Scattered from cluster 6 of the first scattering.

| ID | Top 10 noun phrases | Papers |
|---|---|---|
| 0 | conclusion \| method \| purpose \| journal \| author \| manuscript \| article \| quality \| background \| editor | 387 |
| 1 | proposal \| paper \| referee \| reliability \| example \| order \| peer review \| evaluation \| science \| application | 270 |
| 2 | nih \| health \| funding \| grant application \| national institute \| grant \| application \| medical research council \| cost \| grant proposal | 104 |
| 3 | ecology \| peer review system \| concern \| ecologist \| model \| simulation \| publication process \| researcher \| system \| evolution | 104 |
| 4 | scientific article \| megajournal \| traditional peer review \| transparency \| plos \| oamj \| oamjs \| scientific soundness \| scientific community \| open access | 67 |

Table 2.3: Top 5 papers for cluster 1 in the second scattering. The papers are ranked by number of citations. The citation counts were obtained from the citation network of the initial set of papers.

| Rank | Title | Cit. | Year | Source |
|---|---|---|---|---|
| 1 | Scientific Peer Review | 108 | 2011 | ANNUAL REVIEW OF INFORMATION SCIENCE AND TECHNOLOGY |
| 2 | Bias in peer review | 79 | 2013 | JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY |
| 3 | Improving the peer-review process for grant applications – Reliability, validity, bias, and generalizability | 72 | 2008 | AMERICAN PSYCHOLOGIST |
| 4 | Selection of research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of Board of Trustees' decisions | 58 | 2005 | SCIENTOMETRICS |
| 5 | Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by Angewandte Chemie International Edition, or rejected but published elsewhere | 48 | 2008 | JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY |

Table 2.4: Top 5 papers for cluster 2 in the second scattering. The papers are ranked by number of citations. The citation counts were obtained from the citation network of the initial set of papers.

| Rank | Title | Cit. | Year | Source |
|---|---|---|---|---|
| 1 | Big Science vs. Little Science: How Scientific Impact Scales with Funding | 31 | 2013 | PLOS ONE |
| 2 | Peer review for improving the quality of grant applications | 23 | 2007 | COCHRANE DATABASE OF SYSTEMATIC REVIEWS |
| 3 | Percentile Ranking and Citation Impact of a Large Cohort of National Heart, Lung, and Blood Institute-Funded Cardiovascular R01 Grants | 20 | 2014 | CIRCULATION RESEARCH |
| 4 | Peering at peer review revealed high degree of chance associated with funding of grant applications | 18 | 2006 | JOURNAL OF CLINICAL EPIDEMIOLOGY |
| 5 | Big names or big ideas: Do peer-review panels select the best science proposals? | 17 | 2015 | SCIENCE |

Table 2.5: Top 5 papers for cluster 4 in the second scattering. The papers are ranked by number of citations. The citation counts were obtained from the citation network of the initial set of papers.

| Rank | Title | Cit. | Year | Source |
|---|---|---|---|---|
| 1 | Deep impact: unintended consequences of journal rank | 23 | 2013 | FRONTIERS IN HUMAN NEUROSCIENCE |
| 2 | Alternatives to peer review: novel approaches for research evaluation | 12 | 2011 | FRONTIERS IN COMPUTATIONAL NEUROSCIENCE |
| 3 | Journal acceptance rates: A cross-disciplinary analysis of variability and relationships with journal measures | 11 | 2013 | JOURNAL OF INFORMETRICS |
| 4 | Open evaluation: a vision for entirely transparent post-publication peer review and rating for science | 11 | 2012 | FRONTIERS IN COMPUTATIONAL NEUROSCIENCE |
| 5 | Toward a new model of scientific publishing: discussion and a proposal | 10 | 2011 | FRONTIERS IN COMPUTATIONAL NEUROSCIENCE |

### 2.4.3 Visualizing the clusters

We visualize clusters using a packed bubble chart. We developed an algorithm to create these charts (see below). The input of our algorithm is an undirected network. In this network, nodes represent clusters of papers, the weight of a node indicates the number of papers in a cluster, and the weight of an edge between two nodes indicates the relatedness of two clusters in terms of citation links.

#### 2.4.3.1 Bubble chart algorithm

Our bubble chart algorithm determines the coordinates of the bubbles, where each bubble is a node in a network. The objective of our bubble chart algorithm is to obtain a visualization in which the bubbles do not overlap, the empty space is minimized, and the positions of the nodes relative to each other reflect their relatedness as accurately as possible. We base our algorithm on the VOS layout algorithm [119] used in the VOSviewer software, but we make modifications in order to avoid overlapping bubbles and to minimize the empty space.

The area of a node is proportional to the weight of the node. Therefore, the radius of a node is the square root of $w$, where $w$ is the weight of the node. Nodes connected by edges with a high weight should be close together. To achieve this, we minimize a weighted sum of the squared Euclidean distances between all pairs of nodes, which is similar to the VOS layout algorithm [119]. The weighting considers the weight of the edges between pairs of nodes. This weighted sum can be understood as the stress $V$ of the network layout, and our objective is to minimize this stress. Mathematically, the stress function $V$ is given by

$$V(x_1, \ldots, x_n) = \sum_{i<j} s_{ij} \|x_i - x_j\|^2 \tag{2.1}$$

where $x_i$ denotes the coordinates of node $i$ in a two-dimensional space, $\| * \|$ is the Euclidean norm, and $s_{ij}$ is the weight of the edge between nodes $i$ and $j$. To avoid overlapping nodes, we add for all pairs on nodes $i$ and $j$ the constraint

$$\|\mathbf{x}_i - \mathbf{x}_j\| \geq r_i + r_j \tag{2.2}$$

where $r_i$ is the radius of node $i$. Minimization of the stress function in Equation 2.1 subject to the constraint in Equation 2.2 is not straightforward, so we developed a minimization algorithm for it.

#### 2.4.3.2 Minimization algorithm

The best strategy to minimize Equation 2.1 while satisfying Equation 2.2 in a network of two nodes (nodes 1 and 2) is to place the nodes adjacent to each other. When we fix the coordinates of node 1, the coordinates where node 2 can be placed form a circle $c(1,2)$ around node 1 (Figure 2.3A). This circle has a radius equal to the sum of the radius of node 1 and the radius of node 2. Now, we also fix the coordinates of node 2 and add node 3 to the network layout. We can use the same strategy to get its coordinates. The adjacent coordinates for node 3 form the circles $c(1,3)$ and $c(2,3)$ (Figure 2.3B). Therefore, the available coordinates to place node 3 are the intersection points of $c(1,3)$ and $c(2,3)$ (Figure 2.3C).

When we add node 4 to the network layout, the available coordinates for this node are no longer all the intersection points of the circles $c(i,j)$, because some coordinates would cause nodes to overlap (Figure 2.3D). Of the available coordinates, we select the ones that result in the lowest stress. We can find these coordinates by calculating the weighted sum of the squared Euclidean distances between node 4 and each node that has already been assigned to coordinates. We proceed in the same way for all other nodes.

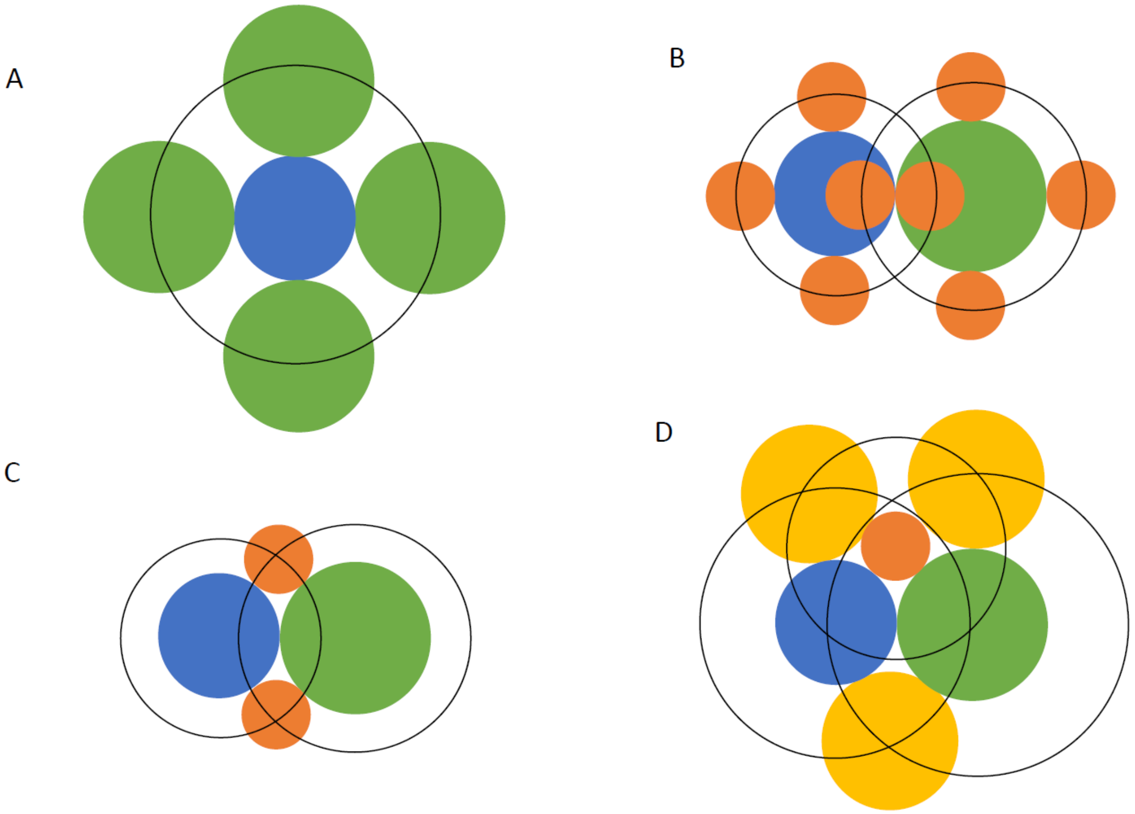


Figure 2.3: Illustration of the minimization algorithm. A: The coordinates for node 2 (green) form a circle around node 1 (blue). B: The coordinates for node 3 (orange) form a circle around node 1 (blue) and another circle around node 2 (green). C: The available coordinates for node 3 (orange) are given by the intersection of the circles in B. D: The available coordinates for node 4 (yellow) no longer include all the intersection points of the circles.

Our minimization algorithm obtains the coordinates of the nodes by adding them one-by-one to the network layout. However, we found that the value of the stress at the end of an algorithm run is highly dependent on the order in which the nodes had been added. To improve our minimization algorithm, we added a step in which we create several lists of the nodes in a different order. For each list, we run the minimization procedure and in the end we return the network layout with the lowest stress.

We order the nodes in the lists as follows. For each node in the network, we create a list with that node as the first node. The next node in the list is the one that is most strongly related to the nodes already in the list. We repeat this process until all nodes have been added to the list.

Our minimization algorithm is a heuristic approach to the minimization of Equation 2.1 and does not guarantee that the global minimum of Equation 2.1 will be found. The pseudocode of the algorithm is provided in the appendix.

## 2.5 Conclusion

We have proposed a tool for scientific literature search based on the Scatter/Gather approach. The tool visualizes the structure of the search results using a packed bubble chart. We have presented a case study demonstrating the use of the tool and we have provided a technical specification of the algorithms included in the tool, in particular the algorithm for creating packed bubble charts.

Compared to traditional literature search tools that present the search results as a list of documents (e.g. Google Scholar), we expect the advantage of our tool to be in the emphasis it puts on showing the structure of the search results. We expect this to be important especially when users are searching not for one specific paper but for a larger set of papers offering a broad understanding of a certain scientific domain. In future work, we plan to test the performance of the tool for different information retrieval tasks.

## 2.6 Data availability

We made available a graphical user interface prototype of the tool, which we named SciMacro (for Science Macroscope) [15].

## 2.7 CRediT author statement

**Juan Pablo Bascur:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.
**Nees Jan van Eck:** Conceptualization, Methodology, Supervision, Writing – review & editing.
**Ludo Waltman:** Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

## 2.8 Appendix

```
-----
INPUT: list INLIST containing nodes ($x_0$,...,$x_n$).
    Each node possesses:
    A node identity $id(x)$
    A radius $r(x)$
    A list of edges $E(x)$ containing ($e_0$,...,$e_n$), with each edge $e$ possessing a weight
    $w(e)$ and a node identity $id(e)$ of the node it connects to
    A coordinate $c(x)$ that contains nothing
OUTPUT: list OUTLIST containing nodes ($x_0$,...,$x_n$) possessing non-empty coordinates $c(x)$
-----
Create list MASTERLIST containing nothing
For each node $x_i$ in list INLIST ($x_0$,...,$x_n$):
    Complete subroutine $S\_ORDER(x_i, (x_0, \ldots, x_n))$
    Create list $Z_i$ containing nothing
    Set coordinate $c(x_{i0})$ of node $x_{i0}$ as $(0, 0)$
    Append node $x_{i0}$ to list $Z_i$
    Set coordinate $c(x_{i1})$ of node $x_{i1}$ as $((r(x_{i0}) + r(x_{i1}), 0)$
    Append node $c(x_{i1})$ to list $Z_i$
    Complete subroutine $S\_COOR(Z_i, (x_{i2}, \ldots, x_{in}))$
    Append list $Z_i$ to list MASTERLIST
Return list OUTLIST in MASTERLIST ($Z_0$,...,$Z_n$), where OUTLIST is the list with lowest
graph stress $V$ as defined in the equation 2.1 $V(OUTLIST)$
------
Subroutine $S\_ORDER$ creates an order of nodes

$S\_ORDER(x_i, (x_0, \ldots, x_n))$:
Create list $X_i$ containing nothing
Append node $x_i$ to list $X_i$ as node $x_{i0}$
Create list $Y_i$ containing nodes $(x_0, \ldots, x_n)$
Remove node $x_i$ from list $Y_i$
While list $Y_i$ containing something:
    For each node $x_j$ in $Y_i$:
        Declare $tw_j$ is the total weight from $x_j$ to all the nodes in $X_i$
    Declare $x_{tw}$ is the node with greatest $tw_j$
    Append node $x_{tw}$ to list $X_i$ as node $x_{ij}$
    Remove node $x_{tw}$ from list $Y_i$
-----
Subroutine $S\_COOR$ gets the coordinates of the nodes for nodes $x_{>1}$

$S\_COOR(Z_i, (x_{i2}, \ldots, x_{in}))$:
For each node $x_{ij}$ in $(x_{i2}, \ldots, x_{in})$:
    Create empty list $TEMP_{ij}$
    For each order-independent pair of nodes $(x_{ijm,}x_{ijn})$ in list $Z_i$, where $m > n$:
        Complete subroutine $S\_TEST(x_{ij}, x_{ijm}, x_{ijn}, Z_i, TEMP_{ij})$
    Append node $temp_{ij}$ to list $Z_i$, where $temp_{ij}$ is the temporal node with lowest node
    stress $v$ in list $TEMP_{ij}$
-----
Subroutine $S\_TEST$ tests if the node $x_{ij}$ can be adjacent to nodes $(x_{ijm}, x_{ijn})$, get the
coordinates of center of these adjacent positions, test if the node $x_{ij}$ on that coordinates
overlaps with other nodes and get the stress of the node $x_{ij}$ on that coordinates.

$S\_TEST(x_{ij}, x_{ijm}, x_{ijn}, Z_i, TEMP_{ij})$:
Declare temporary node $temp_{ijm}$ with coordinate $c(x_{ijm})$ and radius $(r(x_{ij}) + r(x_{ijm}))$
Declare temporary node $temp_{ijn}$ with coordinate $c(x_{ijn})$ and radius $(r(x_{ij}) + r(x_{ijn}))$
If $temp_{ijm}$ and $temp_{ijn}$ DO overlap:
    Declare coordinates $coor_{ijmn1}$ and $coor_{ijmn2}$ are the coordinates of the intersection
    between the borders of $temp_{ijm}$ and $temp_{ijn}$
    For $coor_{ijmnk}$ in list $(coor_{ijmn1}, coor_{ijmn2})$:
        Declare temporary node $temp_{ijmnk}$ is a node with the parameters of node $x_{ij}$, except
        that its coordinate $c(temp_{ijmnk})$ is $coor_{ijmnk}$
        If node $temp_{ijmnk}$ DOES NOT overlaps with any node in $Z_i$:
            Declare node stress $v_{ijmnk}$ is the total stress of the node $temp_{ijmnk}$ with
            every node in the list $Z_i$
        Append $temp_{ijmnk}$ to list $TEMP_{ij}$
```

-----

# Chapter 3

# Academic information retrieval using citation clusters: In-depth evaluation based on systematic reviews

## Abstract[1]

The field of science mapping has shown the power of citation-based clusters for literature analysis, yet this technique has barely been used for information retrieval tasks. This work evaluates the performance of citation-based clusters for information retrieval tasks. We simulated a search process with a tree hierarchy of clusters and a cluster selection algorithm. We evaluated the task of finding the relevant documents for 25 systematic reviews. Our evaluation considered several trade-offs between recall and precision for the cluster selection. We also replicated the Boolean queries self-reported by the systematic reviews to serve as a reference. We found that citation-based clusters' search performance is highly variable and unpredictable, that the clusters work best for users that prefer recall over precision at a ratio between 2 and 8, and that the clusters are able to complement query-based search by finding additional relevant documents.

## 3.1 Introduction

Researchers and other knowledge workers need special information retrieval (IR) tools because their IR tasks and practices differ from the general public and from each other [55, 100, 136]. Academic literature search is an essential part of any research project, and the most commonly used IR method is query-based retrieval: search using keyword queries to retrieve a ranked list of documents. However, some users complement this method with citation-based IR methods that follow the citations of the documents [79, 124]. These methods have two major advantages over query-based retrieval: 1) They are independent of the keywords, helping with lack of vocabulary knowledge or semantic ambiguity, and 2) they use the intellectual information of the citations, helping find documents that other researchers already connected. However, these methods can be timewise inefficient for users [175].

Given the prominence of citation clusters in scientometric research [164], it is remarkable that citation cluster-based IR (CCIR) is largely absent from the toolset of users [173]. CCIR combines

[1] This chapter is based on: Juan Pablo Bascur, Suzan Verberne, Nees Jan van Eck and Ludo Waltman. 2023. Academic information retrieval using citation clusters: In-depth evaluation based on systematic reviews. Scientometrics, 128, 2895–2921. https://doi.org/10.1007/s11192-023-04681-x [17]

citation-based IR and cluster-based IR by making use of clusters of documents identified based on citation links. CCIR could allow users to also use approaches developed in scientometric research, such as science maps [38], cluster labeling [144], and visualization software [156]. CCIR offers two potential benefits over other citation-based IR methods: 1) it is less hindered by documents that cite the relevant literature poorly [134] and 2) it communicates the topic structure of a document corpus, including the relative size of different topics and the relations between topics [129].

Effective cluster-based IR requires the clusters to group together the documents that are relevant for the IR task of the user (i.e., the cluster hypothesis [160]). The extent to which this condition is fulfilled by CCIR is an open question. The answer may be different for different types of IR tasks [73] and for different CCIR implementations. We consider one specific IR task, namely performing a literature search to write a systematic review (SR), and one specific CCIR implementation, namely a tree hierarchy of citation-based clusters of MEDLINE documents. As discussed below, we believe this to be a sensible use of CCIR. Moreover, data for experimentation was relatively easily available for this task. To determine the extent to which CCIR groups together relevant documents, we address the following research questions:

- What types of users are best served by CCIR?
- What types of SRs are best served by CCIR?
- What are the strengths and weaknesses of CCIR?

We answer these questions by simulating a CCIR search process, evaluating its performance and analyzing its results. We simulated the CCIR search process in the tree hierarchy with an algorithm that aims to simulate the behavior of a human user. The idea of a CCIR hierarchy is based on classical cluster-based IR strategies [48, 92] and on a frequently used scientometric approach for creating classification systems of science [164]. We evaluated the performance of CCIR for the task of finding the relevant documents for 25 SRs from a benchmark dataset [139], using as performance reference the SRs' self-reported Boolean query search retrieved documents, obtained through intensive manual annotation. This task is well-suited for cluster-based IR because all relevant documents are considered equally important; the task is considered a Boolean retrieval task, so there is no ranking of documents. From these results we analyzed the different preferences of hypothetical users regarding the trade-off between precision and recall, the overlap between documents retrieved by CCIR and by a Boolean query, and how the topic of a SR affects its task performance.

To our knowledge, our work is the first study that evaluates the performance of CCIR. We additionally provide two outputs that can be reused by other researchers: 1) an evaluation protocol for clusters-based IR methods that uses SRs, and 2) an extension of the original SR dataset with the annotated Boolean queries.

This paper is organized as follows. We discuss related work in Section 3.2, explain out methodology in Section 3.3, show our results in Section 3.4, discuss our results in Section 3.5, and conclude our work in Section 3.6.

## 3.2 Related work

### 3.2.1 Science mapping

Our research on CCIR is part of a bigger trend of research that attempts to connect the fields of scientometrics and information retrieval. Experts agree that these fields have much to gain from each other [63, 112]. While research on CCIR seems to have slowed down in recent years, research on clustering methods in the field of scientometrics continues to move forward.

Closest to our research are the citation clusters used for science mapping and field delineation studies [38, 42]. It has been shown that these clusters create communities of documents with semantic similarity (i.e., a common topic) [97] and that they provide insights for analyzing these documents [146]. Citation clusters are also used to represent communities of documents in the visualization of a citation network (which is a network of documents and their citations to each other) [37, 158].

Text similarity-based clusters, both on their own [31] and enriched with citations [4, 91], have also been used to map science. Waltman et al. [165] compare citation-based similarity clusters with text similarity-based clusters. We decided not to include the use of text similarity in our research because text similarity-based cluster IR is already a well-studied method (see Section 3.2.3).

### 3.2.2 Citation-based IR

Citation-based IR methods are frequently used in academic search. The most common method is to retrieve the documents that cite or are cited by a given document (a.k.a. citation tracking). A further step of this method is to track the citations of these retrieved documents (a.k.a. snowballing). Some of the developments in citation-based IR are tools to track citations [87, 90, 99, 105, 107, 157, 163], protocols to find relevant documents to write a SR by tracking citations [20, 85], tools that delineate fields by tracking citations [184], methods to rank search results by tracking citations [21, 116], and methods to find the seminal documents of a topic by tracking citations [68]. Additionally, citation-based IR is addressed by the communities around the workshop series Bibliometric-enhanced Information Retrieval (BIR) [63] and the related workshop series Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) [30].

The most significant difference between CCIR and citation tracking is that CCIR creates clusters and retrieves documents using the structure of the whole citation network, while citation tracking retrieves documents using only the structure of the documents closest to the initially selected document in the citation network. Both methods focus on different aspects of the citation network, so both can be valuable to the academic IR toolset.

### 3.2.3 Cluster-based IR

Cluster-based IR methods retrieve one or more clusters of documents, and these clusters are usually based on text similarity. These methods have been used for academic search both in commercial context [88] and academic contexts [122], and have also included the text from cited documents in their similarity score [1]. Non-academic IR has also been used to cluster web search results [147]. Additionally, the seminal Scatter/Gather browsing model [48] (on which we draw inspiration for our evaluation) proposes a user interaction protocol where the user removes irrelevant documents over several iterations by creating new sets of documents using the clusters from the previous iteration. Bascur et al. [16] proposed specifications for a CCIR tool that uses the Scatter/Gather model.

Cluster-based IR works have a wide methodological variety, reflected in the following methodological choices:

- Relatedness attribute between documents: Connections (e.g. citations, as we did) or shared elements (e.g. text, authors, keywords);
- Which set of documents to cluster: Either the whole corpus (as we did) or a subset of the corpus that is retrieved by a query;
- What is the structure of the clustering solution: Either hierarchical (as we did) or flat (a.k.a. independent clusters);
- How to select clusters during the evaluation: Either select clusters using knowledge of the document relevance (as we did) or select clusters using a query match;
- How to retrieve documents during the evaluation: Either retrieve all documents within a cluster (as we did) or retrieve only some.

Our purpose is not to compare the pros and cons of each of these methodological choices. Instead, our focus is on evaluating the specific methodological choices considered in our work. Similar to our work is the work of He et al. [71], who visualize academic search results using, among other elements, citation-based clusters. The difference between their approach and ours is that we use the clusters as a means to retrieve documents, while they use the clusters for visualization of search results. In their work, they showed that their visualization can increase the efficiency (i.e., completion time) and user satisfaction for complex tasks, but not for simple tasks. This result suggests that the effectiveness of CCIR may depend on the task. Therefore, we look at individual SR tasks to see how the effectiveness differs between them.

Measuring the effectiveness of clustering, both for IR and for other purposes, is not trivial, as no clustering solution can satisfy every possible search task [181]. Our approach is to measure clustering effectiveness without the participation of real users (a.k.a. offline evaluation). Many other studies have adopted the same approach. For instance, Abdelhaq et al. [67] created a metric for evaluating Twitter data clustering based on the stability and coverage of the most common keywords in a cluster. In a bioinformatics example, Atkinson et al. [82] evaluated the effectiveness of a gene similarity network clustering by observing to what extent each cluster had a single gene function. Yuan et al. [181] created novel metrics that consider the number of clusters necessary to retrieve a given percentage of the relevant documents. De Vries et al. [49] created an evaluation framework where the relevant documents are known and the clustering solution is compared with a random baseline. Abbasi and Frommholz [1] evaluated clustering with a simulation where a virtual user already knows which are the relevant documents. Our evaluation is most similar to the latter two studies because our cluster selection algorithm already knows which are the relevant documents, which is a common assumption in evaluation of retrieval methods [110].

## 3.3 Method

### 3.3.1 Task design and data collection

The task we address is to find the documents necessary to write a given SR. The data that we use for this task comes from the dataset published by Scells et al. [139] (from now on referred to as the Scells dataset). This dataset contains:

- 177 SRs published by the Cochrane library between 2014 and 2016.
- The references of each SR that belong to the included studies or excluded studies category of that SR. We consider both categories necessary for the task of writing a SR, so we included documents from both categories in the set of relevant documents of the task (see below for an explanation).
- The self-reported Boolean query that the authors of each SR used when they searched using the OVID search platform with the MEDLINE database, hereafter referred to as the Boolean query.

We intend to retrieve the documents that the authors of the SR found in their search, thus we use the authors' Boolean queries to retrieve documents. We retrieved these documents following these steps:

1. We manually confirmed that the Boolean queries in the Scells dataset were the same as the ones self-reported by the SRs, and when this was not the case, we used the self-reported one.

2. We translated the Boolean queries from the OVID format into the PubMed format because the OVID search platform does not have an API service, while the PubMed search platform does [138] and it also includes the MEDLINE database. We translated the formats using the TRANSMUTE software [140] and then we manually checked that the translation was correct (i.e., that both formats would retrieve the same documents). Some translations were not possible because the OVID search platform provides functionalities that the PubMed search platform does not (e.g., word distance-based arguments). A full report on the translations and how we handled difficult cases can be found in the supplementary material, Tables S1 and S2.

3. For each SR, we performed a search using the PubMed API based on the PubMed Boolean query, and we included the retrieved documents in the document set retrieved by the Boolean query.

4. We removed from the retrieved document set the documents that were not in the citation network (which is described in Section 3.3.2). We also removed from the relevant document set (see below) the documents that were absent from the document set retrieved by the Boolean query in order to maintain consistency between both sets (i.e., so that the relevant document set is a subset of the document set retrieved by the Boolean query).

To improve the quality of our evaluation, we selected a subset of the SRs in the Scells dataset to be used in our evaluation. Our selection criteria were:

- The relevant document set contains at least 10 documents. We chose this value because with fewer relevant documents, the increase in recall for each retrieved document would be more than 0.1 and we wish a more fine-grained increase to facilitate interpretation of the results.
- The number of retrieved documents self-reported by the authors (i.e., from all their search sources) is of a similar order of magnitude (i.e., between 10 times less and 10 times more) as the size of the document set retrieved by us with the Boolean query. This condition excludes SRs whose self-reported number of retrieved documents is vastly different from ours.

This selection resulted in 25 SRs (see Figure 3.4A in Section 3.4 for the number of relevant documents per SR), of which 7 were published in 2014, 10 in 2015 and 8 in 2016. The number of SRs may seem small, for instance in comparison with the work by Janssens et al. [90], who used 250 SRs. However, we manually annotated the Boolean queries, which is very labor intensive. Additionally, while the number of SRs is modest, the number of document in our citation networks is very large ($\tilde{7}$ million per network, see below).

Cochrane library SRs have, for our purposes, three categories of documents in their references:

- Included studies: Studies that provide information that advances the objective of the SR.
- Excluded studies: Studies that were considered for the included studies category but were discarded because they did not match the selection criteria of the SR.
- Additional references: Documents that were not considered for the included studies category.

The Cochrane library has a clear rule for which documents should go into the excluded studies category: When a user discards a document, after they have read the document full text to any extent, the document is an excluded study, else it is not (e.g., discarded after reading the abstract).

We decided to regard the excluded studies as relevant documents for the retrieval task because, by the above rule, the user needs to find and read these documents in order to exclude them. Additionally, the selection criteria that discard an excluded study can be so particular (e.g., number of participants in the study) that we believe it is not reasonable to expect an IR tool to be able to discard these documents.

### 3.3.2 Citation network

We needed to create a citation network for the tree hierarchy of clusters. We used the in-house Dimensions database, which contains all the documents included in MEDLINE and also their citation links. We created the citation network following these steps:

1. We retrieved all the documents contained in the Dimensions database.

2. We removed all the documents published the same year or later than the SRs to make sure we do not provide unfair advantageous information to the clustering (see below). Therefore, we created a different citation network for each publication year in the Scells dataset: One until 2013, another until 2014 and another until 2015.

3. We limited the documents of the citation networks to the ones available in the MEDLINE database, because the self-reported Boolean queries were performed exclusively within the MEDLINE database. We identified the MEDLINE documents using the PubMed database available at Leiden University's Centre for Science and Technology Studies (CWTS).

4. Because of the computing resources needed to handle large citation networks, we limited the publishing years of each network to 11 years (2003-2013, 2004-2014, and 2005-2015).

The sizes of the citation networks were:

- Citation network 2003-2013: 6,549,426 documents, 81,284,099 citation links.
- Citation network 2004-2014: 6,879,646 documents, 86,001,142 citation links.
- Citation network 2005-2015: 7,194,514 documents, 90,164,417 citation links.

Documents that are in the reference lists of a given SR are connected to the SR by a citation link. These connections help the clustering algorithm to put all these documents in the same cluster, which would artificially increase the performance of CCIR. This is not fair because in a real scenario these connections could not exist because the SR has not been published yet. We removed not only these connections, but all the documents published in the same year and in later years because they could be influenced by these connections. Because we remove the documents published in the same year, we may also remove some documents that existed before the publication of the SR. However, none of the relevant documents were removed in this process.

### 3.3.3 Simulation of CCIR

In this section we explain how we simulated the CCIR search process so we can evaluate the performance of CCIR.

#### 3.3.3.1 Clustering

We created a tree hierarchy of clusters for each citation network. We started by clustering the documents into at most 10 clusters, based on the idea that in practice it may be difficult for users to handle more than 10 clusters. Then, the documents of each cluster were again clustered into at most 10 smaller clusters, and so on. As discussed below, the documents that could not be included in these clusters were excluded from the tree. This process created a nested tree of clusters with a depth of 13 levels (not counting the root level). We only clustered into smaller clusters the clusters that contained relevant documents because otherwise they were irrelevant for the evaluation.

We performed the clustering using a methodology built on the work of Waltman and van Eck [164]. This methodology is used in combination with the Leiden algorithm [153]. This combination provides a state-of-the-art approach for document clustering in the field of scientometrics. This approach has been used in a large number of research articles (e.g. [28, 76, 143]). It is also used in products of the analytics companies Elsevier [57] and Clarivate [130]. We therefore consider it the state-of-the-art approach for citation-based clustering.

In the methodology of Waltman and van Eck [164], the tree hierarchy is built in a bottom-up manner while we take a top-down approach. We made this change because it reflects how a real user would create a tree, going from the general to the specific. It also saves computer resources by not creating sub-clusters for clusters that are of no interest. Another change is that Waltman and van Eck merged small clusters based on a cluster size threshold, while we merged small clusters based on a number of clusters threshold (at most 10 clusters, as mentioned before). We made this change because for a real user it is more intuitive to control the maximum number of clusters than the minimum number of documents per cluster.

The purpose of the Leiden algorithm is to assign documents to clusters based on the connections between the documents. The algorithm rewards pairs of documents in the same cluster that are connected by a citation link and penalizes pairs of documents in the same cluster that are not connected. The magnitude of the penalty is determined by the resolution parameter of the algorithm, which must be provided externally. A higher resolution leads to more and smaller clusters.

Mathematically, the clustering algorithm maximizes the following quality function:

$$V(x_1, \ldots, x_n) = \sum_{i=1} \sum_{j=1} \delta(x_i, x_j)(a_{ij} - r) \tag{3.1}$$

In this quality function, $i$ and $j$ are documents, $x_i$ is the cluster of document $i$, and $r$ is the resolution parameter. $a_{ij}$ equals 1 if there is a citation link between documents $i$ and $j$. Otherwise $a_{ij}$ equals 0. $\delta$ equals 1 if $x_i$ and $x_j$ are equal (i.e., documents $i$ and $j$ are in the same cluster). Otherwise $\delta$ equals 0.

The Leiden algorithm returns the clustering solution that maximizes Equation 3.1. To limit the number of clusters per clustering (i.e., children clusters per parent cluster) to at most 10, we merged the smaller clusters following these steps:

1. If there are more than 10 clusters in the clustering solution, select the smallest cluster. If there is a tie in the size, randomly select one of the smallest clusters. If the number of clusters is 10 or fewer, stop.
2. If there are no citation links between the documents in the selected cluster and documents outside the selected cluster, remove the selected cluster from the clustering solution and then go back to step 1.
3. For each cluster other than the selected cluster, calculate the highest resolution under which this cluster would merge with the selected cluster (method from Waltman and van Eck [164]). This resolution is always lower than the current resolution because otherwise the clustering algorithm would have already merged these clusters.
4. Merge the selected cluster with the cluster for which the highest resolution was obtained in step 3, and then go back to step 1.

The resolution parameter must be provided externally, but the literature has not yet established a rule of thumb for selecting a suitable value (although the work of Sjögårde and Ahlgren [142, 143] goes in that direction). We therefore used our own heuristic. Using a trial-and-error approach, we tried to find resolution values for each level so that the following conditions were satisfied as much as possible:

- The size of the 10 largest clusters after merging was similar to the size of these clusters before merging. This condition aims to minimize the effect of cluster merging.
- The 10 largest clusters after merging were of similar size. This condition aims to avoid creating one or a few clusters with a disproportionally large number of documents.

Our heuristic resulted in a resolution of $2 \times 10^{-6}$ for the first level of the tree hierarchy. For each subsequent level we multiplied the resolution by 3. At level 13 the resolution is greater than 1 ($2 \times 10^{-6} \times 3^{12} = 1.06$) which is why we have 13 levels (a resolution greater than 1 yields only singleton clusters).

#### 3.3.3.2 Cluster selection

We use a greedy algorithm to select the clusters, starting from the root of the tree hierarchy. The algorithm goes down the tree hierarchy selecting child clusters based on their score, until none of the child clusters has a score higher than the currently selected cluster (see Figure 3.1). We use a greedy algorithm because this reflects how a real user would navigate a tree hierarchy. The score function is the F-score of retrieving the documents in a cluster, determined based on the relevant documents of a given SR:

$$F_\beta = (1 + \beta^2) \times \frac{\text{precision} \times \text{recall}}{(\beta^2 \times \text{precision}) + \text{recall}} \tag{3.2}$$

The precision and recall of each cluster are calculated based on the number of documents in the cluster (i.e., number of positives), the number of relevant documents in the cluster (i.e., number of true positives), and the number of relevant documents not in the cluster (i.e., number of false negatives). A real user does not have access to these numbers. The greedy algorithm therefore simulates an optimistic scenario in which a user is able to accurately assess the quality of different clusters.

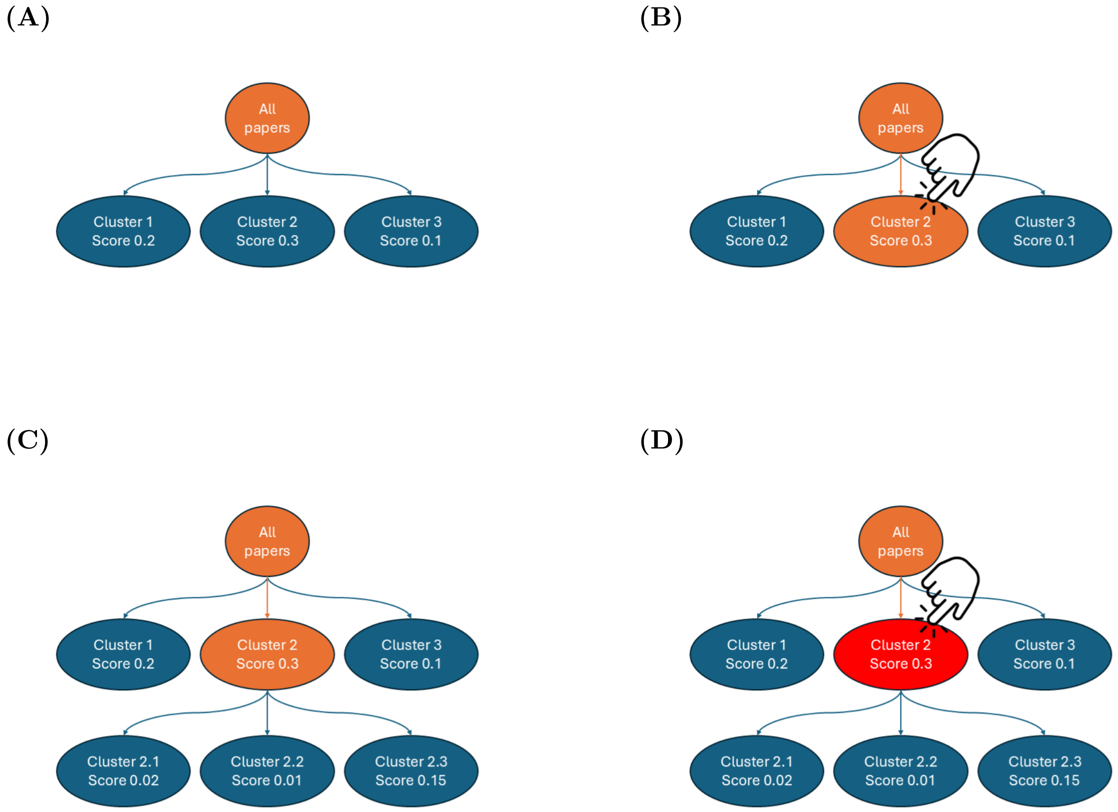


Figure 3.1: Cluster selection algorithm. The bubbles represent clusters of documents. The text in a bubble shows the label and the score of a cluster. The lines are the connections between the parent and the child clusters in the tree hierarchy. The arrows point toward the child clusters. Only the child clusters of the selected clusters are shown. The orange bubbles represent the clusters selected by the algorithm. The orange lines indicate the path followed by the algorithm. The pointer finger shows the selection of the algorithm. **A**: Calculate the score of each cluster at the highest level of the tree hierarchy (Clusters 1, 2, and 3). **B**: Select the cluster with the highest score (Cluster 2). **C**. Calculate the score of each child cluster of the selected cluster (Clusters 2.1, 2.2, and 2.3). **D**. Retrieve the cluster that was already selected (Cluster 2) because it has a higher score than any of the child clusters.

The parameter $\beta$ of the F-score function (Equation 3.2) reflects how a hypothetical user balances recall against precision [160]: Lower values of $\beta$ favor precision, while higher values favor recall. If $\beta = 1$, precision and recall have equal weight. For each SR we retrieve several clusters, each one using different values of $\beta$ to cover a wide range of precision-recall trade-offs: $\beta \in \{0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128\}$. The idea of using a greedy algorithm and different values of $\beta$ to reflect real users is inspired by the “what-if” experiments methodology [10].

### 3.3.4 Quantitative analysis

For our quantitative evaluation, we group the results of the SRs according to value of $\beta$ used by the cluster selection algorithm. In this way, we can compare the aggregated results for different values of $\beta$. We report the number of retrieved documents, the tree-level of the retrieved cluster, precision, recall, and F-score ($\beta = \beta$ used by the cluster selection algorithm).

We report four more metrics that are generated by comparing the cluster selection algorithm results with the Boolean query retrieved documents:

- Intersection proportion of the cluster selection algorithm: Proportion of the documents retrieved by the cluster selection algorithm that are also retrieved by the Boolean query.
- Intersection proportion of the Boolean query: Proportion of the documents retrieved by the Boolean query that are also retrieved by the cluster selection algorithm.
- Ratio of retrieved documents: Number of documents retrieved by the cluster selection algorithm divided by the number documents retrieved by the Boolean query.
- F-score difference: F-score of the cluster selection algorithm minus the F-score of the Boolean query ($\beta = \beta$ used by the cluster selection algorithm).

The purpose of the F-score difference is to evaluate the performance of CCIR while also taking into consideration the difficulty of the task for the authors of the SR. We refrain from using the F-score difference to make claims about the relative performance of CCIR compared to the Boolean query. We do not consider such claims to be justified, because there are too many issues that we are not able to take into account in our analyses. For instance, we assume that the Boolean query retrieves all relevant documents, but we are unable to assess the accuracy of this assumption. Also, in practice, a Boolean query is written over several iterations of trial and error. We are unable to analyze the impact of this iterative process, since we have access only to the final version of a Boolean query.

Instead of directly comparing the performance of a CCIR approach with a Boolean query approach, our quantitative analysis focuses on answering the following questions:

- To what extent does the performance of CCIR varies between individual SRs? We answer this by analyzing the dispersion of the F-score difference grouped by of $\beta$.
- How similar are the sets of documents retrieved by CCIR and the Boolean query? We answer this by analyzing the intersection proportion of both CCIR and the Boolean query
- For which values of $\beta$ is CCIR more effective? We answer this by analyzing most of the quantitative metrics, and how their values change when the value of $\beta$ increases or decreases.

### 3.3.5 Qualitative analysis

In our qualitative analysis we address the following questions:

- How does the nature of a SR affect the performance of CCIR and a Boolean query?
- What type of documents does CCIR or a Boolean query retrieve or miss?

We address these questions by an expert reading of the SRs performed by the first author of our paper (Juan Pablo Bascur), who is trained in the biomedical field, and supported by an expert in Boolean query searches for biomedical purposes (Jan W. Schoones).

We performed the qualitative analysis on the retrieved documents of three SRs. We selected the SRs based on their F-score difference for $\beta = 4$ (we used $\beta = 4$ because it had the highest recall dispersion, which helps highlight the differences between SRs; see Section 4). We selected the SRs with the lowest, highest and third highest F-score difference, which in the Scells dataset correspond to the ids SR59 [126], SR47 [46] and SR80 [109], respectively.

For each SR, we characterized:

- Goal: The question that the authors of the SR want to answer.
- Needs: The nature of the documents that the authors need to retrieve to achieve the goal.
- Boolean query components: The components of which the Boolean query consist. A component is a group of Boolean terms that belong to the same topic.

For each SR we also selected one of the clusters that CCIR retrieved for this SR, that we subjectively found it had good precision and recall (hereafter known as the optimal cluster). We also selected from the clusters that CCIR retrieved the parent and the child of the optimal cluster to expand the range of our analysis, but we discarded the child clusters because they were so small that they did not provide qualitative information. Therefore, we selected the parent of the optimal cluster, hereafter known as the parent cluster.

We inferred the topic of each set of documents (these are, the clusters and the document retrieved by the Boolean query) from the titles of the documents. For the bigger document sets, we facilitated this process by inferring the topics from the most common noun-phrases in the titles of the documents. We extracted noun phrases from titles using the spaCy Python library [83].

To guide our analysis, we use Venn diagrams of the overlap between the relevant documents, the selected clusters of CCIR and the documents retrieved by the Boolean query. We also look for documents retrieved by CCIR but not by the Boolean query that, given their nature, could have been relevant documents if the authors of the SR had found them.

## 3.4 Results

### 3.4.1 Quantitative results

In this section we describe the quantitative analysis of the 25 SRs evaluation results. Figure 3.2 shows the precision, recall, F-score and F-score difference, Figure 3.3 shows the intersection proportions, Figure 3.4 shows the number and ratio of retrieved documents, and Figure 3.5 shows the level of the selected clusters.

#### 3.4.1.1 To what extent does the performance of CCIR vary between individual SRs?

Figure 3.2E shows that the F-score difference values have a large dispersion: within $\beta$ groups the interquartile range is 0.2 or higher, and the highest range (at $\beta = 4$) is 0.5. This result shows that the performance varies between SRs, and it highlights the importance of analyzing individual SRs in the qualitative analysis presented in Section 3.4.2.

#### 3.4.1.2 How similar are the sets of documents retrieved by CCIR and the Boolean query?

Figure 3.3 shows that these two sets of documents are very different because their intersection proportion is very low. We analyzed Figure 3.3 focusing on three $\beta$ groups, which we selected based on Figure 3.4D: when both document sets are of the same size ($\beta = 16$), when the CCIR set is 10 times bigger than the Boolean query set ($\beta = 128$), and when the CCIR set is 10 times smaller than the Boolean query set ($\beta = 2$). When both sets are the same size and when the CCIR set is 10 times bigger, the intersection proportion is surprisingly low: 0.1 for the former (Figures 3.3A and 3.3B) and 0.5 for the latter (Figure 3.3B). When the CCIR set is 10 times smaller, the proportion is also low (0.6), but additionally this value starts to fall dramatically on the subsequent groups of $\beta$ (Figure 3.3A).

#### 3.4.1.3 For which values of $\beta$ is CCIR more effective?

Figure 3.5 shows that the tree-level of the selected clusters is linearly correlated with the value of $\beta$ (using our powers of 2 scale), or in other words, the median level goes up by 1 level for each sequential $\beta$ value.

Figure 3.4D shows that, for $\beta$ between 2 and 128, the CCIR retrieved document set was between 10 times smaller and 10 times bigger than the Boolean query document set. Figure 3.2B shows that the $\beta$ groups after $\beta = 8$ have less precision than the Boolean query (0.025, Figure 3.2A). Figure 3.2C shows that recall improves little after $\beta = 8$. Therefore, we think that the results of groups

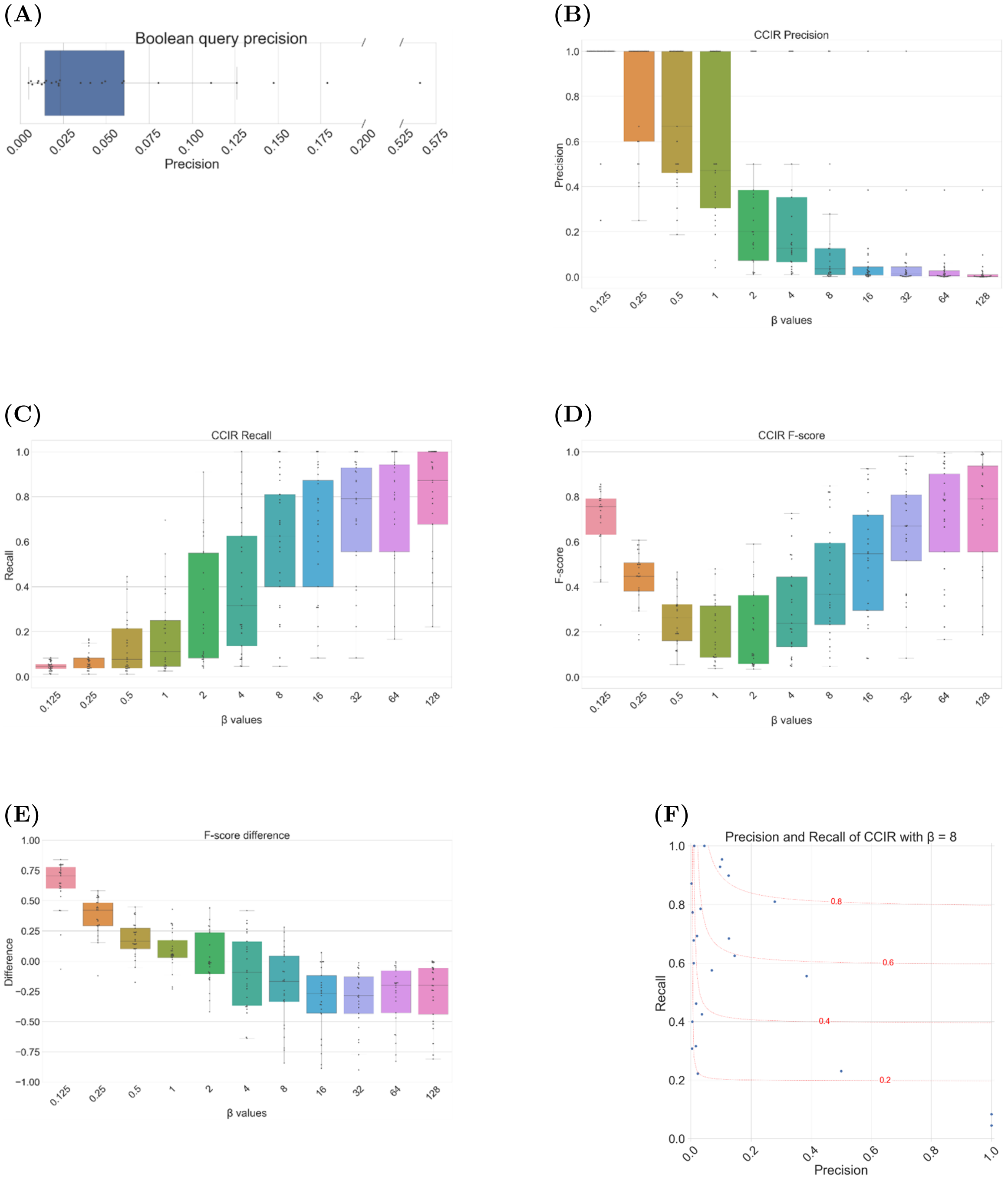


Figure 3.2: Precision, Recall and F-Score. **A**: Precision of the Boolean query. Each data point is a SR, and the X axis is the precision. **B** to **E**: Each data point is a SR, the X axis is the $\beta$ group, and the Y axis is the respective metric of that $\beta$ group for that SR. **B**: Precision of CCIR. **C**: Recall of CCIR. **D**: F-Score of CCIR. **E**: F-score difference between CCIR and the Boolean query (CCIR minus Boolean query). **F:** Precision and recall of $\beta = 8$. Each data point is a SR, the X axis is the precision of CCIR, the Y axis is the recall of CCIR, and the red lines are the isocurves of the F-score ($\beta = 8$).

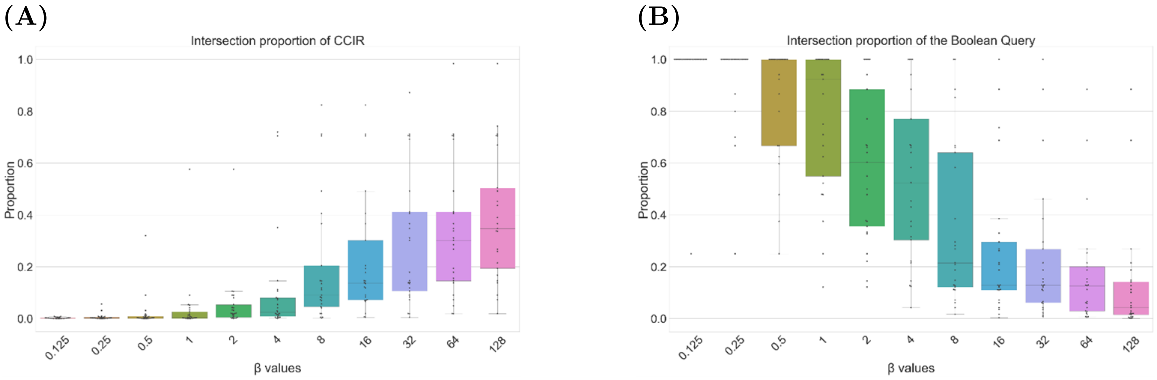


Figure 3.3: Intersection proportions. **A and B**: Each data point is a SR, the X axis is the $\beta$ group, and the Y axis is the respective metric of that $\beta$ group for that SR. **A**: Intersection proportion of CCIR. **B**: Intersection proportion of the Boolean query.

$\beta = 2$, $\beta = 4$ and $\beta = 8$ balance size, precision and recall the best. Also, outside these groups the balance decreases much faster from $\beta = 1$ to the lower values of $\beta$ than from $\beta = 16$ to the higher values of $\beta$.

### 3.4.2 Qualitative results

In this section we describe the qualitative analysis of three selected SRs and their evaluation results. Figure 3.6 shows their Venn diagram of the intersection between the Boolean query, the CCIR and the relevant documents. Table 3.1 shows their quantitative data, Table 3.2 shows their characterization and Table 3.3 shows the topic of their sets of documents. The details on the construction of their Boolean query components can be found in supplementary material Figures S1, S2 and S3, and their topics in supplementary material Tables S3, S4 and S5.

#### 3.4.2.1 SR59: Retinoic acid post consolidation therapy for high-risk neuroblastoma patients treated with autologous hematopoietic stem cell transplantation

This SR had the lowest F-score difference and also a high Boolean query precision (Table 3.1). Its goal was to determine if patients with the condition *Neuroblastoma* recuperate better from the treatments *Chemotherapy* and *Bone Marrow Transplant* if they are treated with the medication *Retinoic Acid* (Table 3.2).

The document set of Boolean query and the two clusters had similar topics, but the cluster topics were missing the component *Retinoic Acid* (Table 3.3), which is one of the needs of SR59 (Table 3.2). This suggests that CCIR did not create a cluster with *Retinoic Acid*, and we wonder why. All the relevant documents of SR59 clearly share a common topic (we read their titles) so it would seem that they should be mostly in the same CCIR cluster. An explanation for this mystery seems to be given by the topic of the parent cluster. Here, we found that the topic fulfills the needs of SR59, except that instead of *Retinoic Acid* it has the component *131L-MIBG*, which is a medication with similar uses to *Retinoic Acid*. It seems then that the existence of a cluster with the needs of SR59 and *Retinoic Acid* was mutually exclusive with the existence of a cluster with the needs of SR59 and *131L-MIBG*, and CCIR created the latter instead of the former because of its higher fitness. This likely resulted in CCIR spreading the relevant documents of SR59 among other clusters, decreasing the F-score difference value.

The Boolean query of SR59 is missing the component *Bone Marrow Transplant* from the needs of SR59 (Table 3.2), yet the Boolean query achieves a high precision (Table 3.1). This is because the

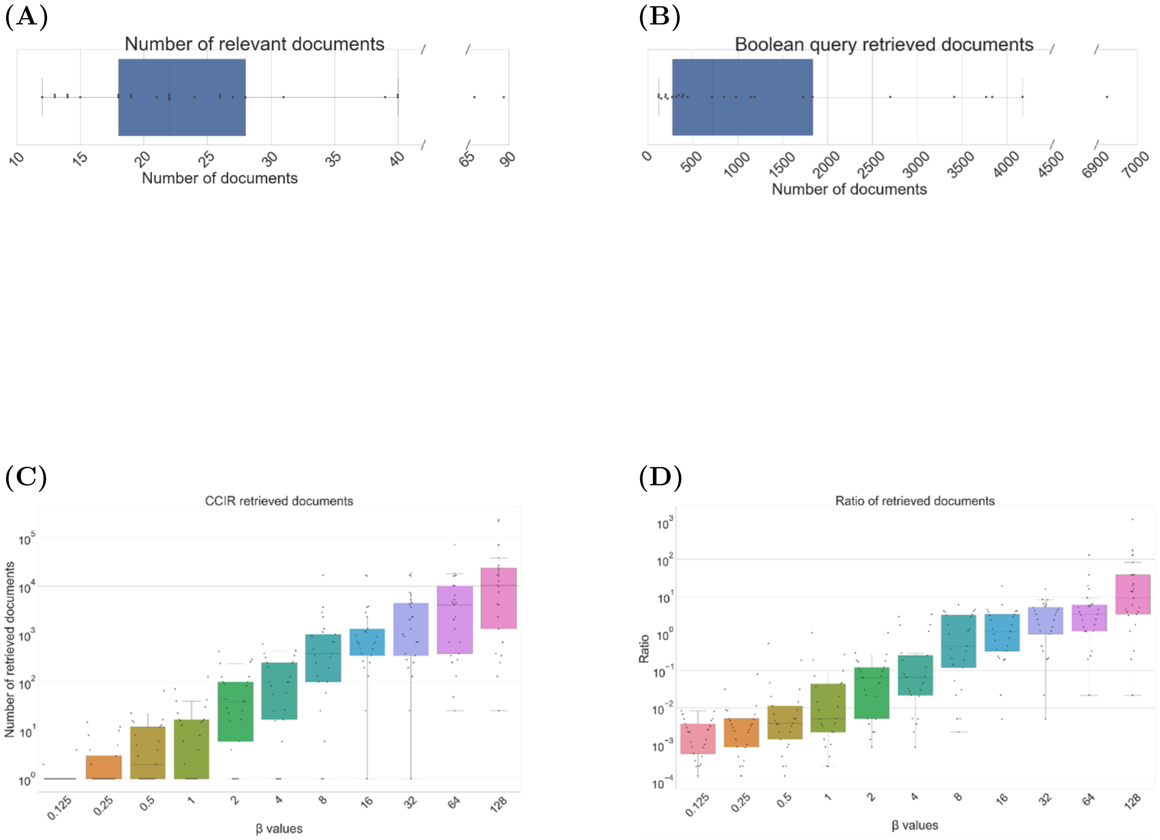


Figure 3.4: Documents sets sizes. **A**: Relevant documents sets sizes. Each data point is a SR, and the X axis is the size of the relevant documents set. **B**: Boolean query retrieved documents sets sizes. Each data point is a SR, and the X axis is the size of the Boolean query retrieved documents set. **C and D:** Each data point is a SR, the X axis is the $\beta$ group, and the Y axis is the respective metric of that $\beta$ group for that SR. **C:** CCIR retrieved documents sets sizes. **D:** Ratio of retrieved documents. Calculated as the CCIR retrieved documents set size divided by the Boolean query retrieved documents set size.

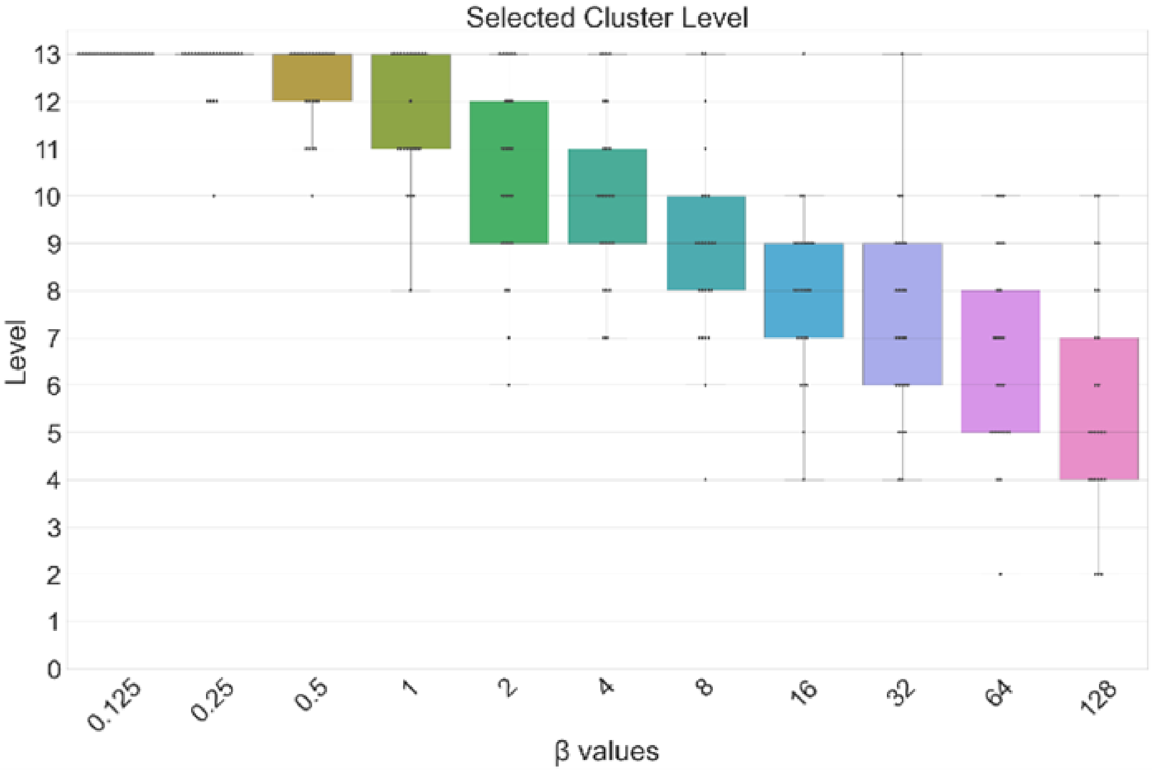
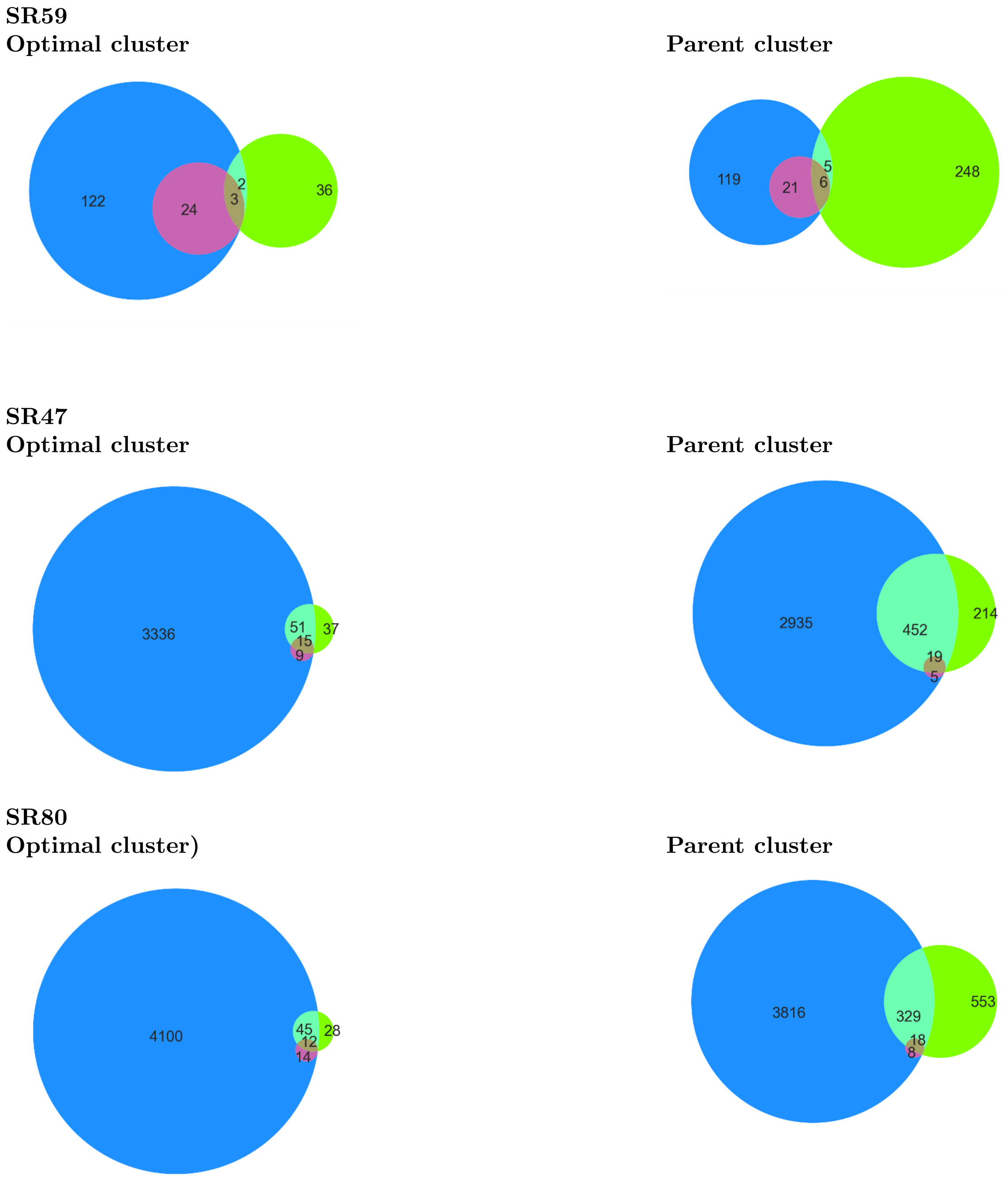


Figure 3.5: Tree-level of the retrieved clusters. Each data point is a SR, the X axis is the $\beta$ group, and the Y axis is the level of the cluster selected by that greedy algorithm for that SR. Level 0 is the set of all documents in the citation network.



SR59

Optimal cluster

Parent cluster

SR47

Optimal cluster

Parent cluster

SR80

Optimal cluster)

Parent cluster

Figure 3.6: Venn diagram of the intersections. Blue: Boolean query retrieved documents set, Green: CCIR retrieved documents set, Red: Relevant documents set.

Table 3.1: Quantitative data of the SRs in the qualitative analysis. These are the SRs selected for qualitative analysis (SR59, SR47 and SR80). The F-score values of $\beta = 4$ were the ones used to select the SRs. The optimal cluster was selected for its good precision and recall, and the parent clusters because it was the parent cluster of the optimal algorithm (see methods, Section 3.3.5).

| **Set of documents** | **Metric** | **SR59** | **SR47** | **SR80** |
|---|---|---|---|---|
| $\beta = 4$ | CCIR F-score | 0.15 | 0.52 | 0.41 |
| | Boolean query F-score | 0.79 | 0.11 | 0.10 |
| | F-scores difference | -0.64 | 0.42 | 0.31 |
| Boolean query | Retrieved documents set size | 151 | 3411 | 4171 |
| | Relevant retrieved documents set size | 27 | 24 | 26 |
| | Precision | 0.18 | 0.01 | 0.01 |
| Optimal cluster | CCIR $\beta$ value | 1 | 2 | 2 |
| | Retrieved documents set size | 41 | 103 | 85 |
| | Relevant retrieved documents set size | 3 | 15 | 12 |
| | Precision | 0.07 | 0.15 | 0.14 |
| | Recall | 0.11 | 0.62 | 0.46 |
| | Intersection set size | 5 | 66 | 57 |
| | Intersection proportion of CCIR | 0.12 | 0.64 | 0.67 |
| Parent cluster | CCIR $\beta$ value | 4 | 16 | 8 |
| | Retrieved documents set size | 259 | 685 | 900 |
| | Relevant retrieved documents set size | 6 | 19 | 18 |
| | Precision | 0.02 | 0.03 | 0.02 |
| | Recall | 0.22 | 0.79 | 0.69 |
| | Intersection set size | 11 | 471 | 347 |
| | Intersection proportion of CCIR | 0.04 | 0.69 | 0.39 |

combination of the components *Neuroblastoma* and *Retinoic Acid* was so infrequent in the literature that it was enough for Boolean query. This shows that the Boolean query can give high precision for highly specific needs.

#### 3.4.2.2 SR47: Surgery for the resolution of symptoms in malignant bowel obstruction in advanced gynaecological and gastrointestinal cancer

This SR had the highest F-score difference (Table 3.1). Its goal was to determine how effective the treatment *Surgery* is to treat the condition *Intestinal Obstruction* when caused by the conditions *Gynecological Cancer* or *Gastrointestinal Cancer* (Table 3.2).

We could not identify the topic of the Boolean query document set because the most common noun-phrases were present in only a minor portion of the documents. This could be either because the set of documents was big and therefore has too much diversity, or because it has several disconnected topics, and we believe the latter explanation is the correct one. On the other hand, the topics of the two clusters (Table 3.3) were similar to the needs of SR47 (Table 3.2).

We believe that the Boolean query has several disconnected topics because the needs SR47 were hard to express in a Boolean query format, which ends up retrieving a noisy set of documents. The needs are documents on *Surgery* to treat *Intestinal Obstruction* due to *Gynaecological and Gastrointestinal Cancer* (Table 3.2). However, the Boolean query cannot specify if *Surgery* treats *Intestinal Obstruction* or treats *Gynaecological and Gastrointestinal Cancer*. This case shows that CCIR can help with searches where the relation between the Boolean query terms is ambiguous.

Additionally, we saw an interesting phenomenon happening with the topics of the clusters. Among their documents, there were three synonym noun-phrases that refer to intestinal obstruction: *Malignant Bowel Obstruction*, *Malignant Colorectal Obstruction* and *Malignant Colonic Obstruction*. The optimal cluster only had the first form, while the parent cluster had all three of them. This

Table 3.2: Characterization of the SRs. These are the SRs selected for qualitative analysis (SR59, SR47 and SR80). Goal: The question that the authors of the SR want to answer. Needs: The nature of the documents that the authors need to retrieve to achieve the goal. Boolean query components: The components of which the Boolean query consist. The details on the construction of the Boolean query components are in the supplementary material, Figures S1, S2 and S3.

| | SR59 | SR47 | SR80 |
|---|---|---|---|
| Title | Retinoic acid post consolidation therapy for high-risk neuroblastoma patients treated with autologous hematopoietic stem cell transplantation | Surgery for the resolution of symptoms in malignant bowel obstruction in advanced gynaecological and gastrointestinal cancer | Rituximab for rheumatoid arthritis (Review) |
| Goal | To determine if **retinoic acid** helps **neuroblastoma** patients recuperate from chemotherapy and bone marrow transplants. | To assess the efficacy of **surgery** for **intestinal obstruction** due to advanced **gynaecological and gastrointestinal cancer**. | To evaluate the benefits and harms of **Rituximab** for the treatment of **Rheumatoid Arthritis**. |
| Needs | **Randomized controlled trials** that evaluate if retinoic acid helps neuroblastoma patients recuperate from bone marrow transplants by **comparing retinoic acid treated patients to untreated patients**. | Documents that mention the **evolution** of patients after **surgeries** to treat **intestinal obstruction** due to advanced **gynaecological and gastrointestinal cancer**. | Studies that compare the outcomes of treatments with **Rituximab** with placebo or other Disease-modifying antirheumatic drugs (**DMARD**). |
| Boolean query components | Retinoic acid **AND** Neuroblastoma **AND** Randomized Controlled Trials and Controlled Clinical Trials | Gynecological or gastrointestinal cancer **AND** Intestinal obstruction **AND** Surgery | Rheumatoid Arthritis **AND** Disease-modifying antirheumatic drugs **AND** Randomized Controlled Trials and Controlled Clinical Trials |

Table 3.3: Topic of the sets of documents of the SRs. These are the SRs selected for qualitative analysis (SR59, SR47 and SR80). We obtained these topics by analyzing the most common noun-phrases in the titles of the retrieved documents. The details on the construction of the topics are in the supplementary material, Tables S3, S4 and S5.

| ID | Set of documents | Topic of the set | Topic of all sets |
|---|---|---|---|
| SR59 | Boolean query | Retinoic Acid for neuroblastoma | Treatments of neuroblastoma |
| | Optimal cluster | Marrow transplant for neuroblastoma. | |
| | Parent cluster | 131I-mibg for neuroblastoma. | |
| SR47 | Boolean query | Disperse topic, most common noun-phrases are too infrequent | Treatments of bowel obstructions in cancer |
| | Optimal cluster | Management of bowel obstructions in cancer, includes non-surgery alternatives | |
| | Parent cluster | More techniques for managing bowel obstructions including emergencies bridge as surgery and self-expandable metal stent. | |
| SR80 | Boolean query | Treat rheumatoid arthritis with several DMARDs | Treatments of rheumatoid arthritis with DMARDs |
| | Optimal cluster | Treat rheumatoid arthritis with few DMARDs | |
| | Parent cluster | Treat rheumatoid arthritis with several DMARDs (including certolizumab pegol) | |

implies that the documents with the first form cite each much more intensely than the documents with the other two forms. We see no science-related reason for this to be the case, so we imagine that this citation pattern arises from a community of researchers with the same writing conventions that cite each other. This citation pattern shows one of the risks of CCIR and of citation-based clustering in general: The citations may not only represent an intellectual relationship between two documents, but also other non-scientific relationships that are of no use for IR purposes.

We saw another interesting phenomenon happening with the topics of the clusters. Two of the most common noun-phrases of the optimal cluster were *Inoperable Bowel Obstruction* and *Octreotide* (which is a medication for inoperable tumors). Both noun-phrases imply that their documents lack surgery, but *Surgery* is a need of SR47. This shows that, even when the F-score difference value is high, CCIR may still not have created a cluster with the topic that the user needs.

#### 3.4.2.3 SR80: Rituximab for rheumatoid arthritis (Review)

This SR had the third highest F-score difference (Table 3.1). Its goal was to evaluate the medication *Rituximab* to treat the condition *Rheumatoid Arthritis.* There are two things we must mention for our analysis of SR80: First, that the medication *Rituximab* belongs to a group of medications called *DMARDs* (which means Disease-Modifying Antirheumatic Drugs), and second, that the needs of SR80 include comparing *Rituximab* treatments with either no treatment (a.k.a. placebo) or other *DMARDs* treatments (Table 3.2).

The topic of the Boolean query and the clusters is the same and fits the needs of SR80. This shows that CCIR created a cluster for the right topic. However, CCIR still missed several relevant documents, which shows that creating a cluster for the right topic can be insufficient. We believe that the reason these relevant documents were not in the clusters is that, even if two documents are about the same topic, they may be poorly connected to each other by direct or indirect citations due to the citing practices of their research community. This result challenges one of the core assumptions of CCIR: That two given documents that share a topic will be directly or indirectly well connected by citations.

It seems that the authors of the SR made the conscious decision of building the Boolean query in such a way that it sacrifices precision in favor of recall. This is suggested by the following difference between the required needs of SR80 and the Boolean query components of SR80 (Table 3.2): SR80 requires comparisons between treatments with *Rituximab* (itself a *DMRAD*) and treatments with placebo or other *DMRADs,* but the Boolean query components do not require a document to mention *Rituximab*, resulting in several retrieved documents that do not serve the needs. We believe that the authors made this decision because they expected many documents that use *Rituximab* to mention it in their metadata under the more general term *DMRADs.* This case shows that CCIR can help with searches where the Boolean query cannot be sufficiently specific.

An interesting observation is that, among the most common noun-phrases, the Boolean query mentions the same *DMRADs* as the parent cluster, but the latter also mentions one extra *DMRAD* (*Certolizumab Pegol*). This is interesting because the component *DMRADs* of the Boolean query searched for all the available *DMARDs*, so it should also have found *Certolizumab Pegol.* We found that this happens because of the MeSH term that the component *DMARDs* uses (*"Antibodies, Monoclonal"[Mesh Terms:noexp]*) does not retrieve *Certolizumab Pegol* (which goes under *"Antibodies, Monoclonal, Humanized"[Mesh Terms:noexp]*). Biologically speaking, *DRMADs* is better described by the latter MeSH term than by the former, but it seems that the convention of the National Library of Medicine is to use the former MeSH term for all *DRMADs* except for *Certolizumab Pegol.* The authors may not have been aware of this because otherwise they presumably would have incorporated the second MeSH term in the Boolean query. We believe that this case shows that CCIR can help Boolean query users to ensure they include all necessary vocabulary in their Boolean query.

We wondered if any of the documents of the parent cluster with *Certolizumab Pegol* in their title may have been a relevant document if the authors of SR80 had seen the document during their literature search. We tested this hypothesis by comparing these documents with the needs SR80.

We found one document [170] which cannot be discarded based only on the title or the abstract, and therefore is a relevant document. This case shows that CCIR can find relevant documents that the Boolean query does not.

## 3.5 Discussion

In this section we discuss our findings in relation to our research questions and then discuss the limitations of our work.

### 3.5.1 What types of users are best served by CCIR?

We can answer questions about users by connecting user preferences for recall and precision with the $\beta$ value (user prefer recall $\beta$ times as much as precision). We saw that $\beta = 2$, $\beta = 4$ and $\beta = 8$ had the best balance, and that outside these $\beta$ values the balance decreases faster for lower $\beta$ values than for higher $\beta$ values. Therefore, we can say that CCIR serves best users that prefer recall over precision with a ratio between 2 and 8 times, and for users outside that range it serves higher ratios better than lower ratios.

We wondered if users that perform a literature search for a SR are within this range of ratios, and we used the Boolean queries values as a proxy to answer this. Figure 3.2A shows that the precision of the Boolean queries is between 0.01 and 0.06, and by definition the Boolean queries have a recall 1.0, so the ratio of recall over precision is 1 over 0.01-0.06, or 17-100, very far from our prior range of 2-8. While it is true that the recall of the Boolean query is unrealistically high, the recall would have to be 10 times lower for the ratio to be within the range, which, given that SR literature searches aim for maximum recall, is unlikely. Therefore, we believe that the users that are best served by CCIR are not users that do a literature search for SR. It is beyond our knowledge which type of user might prefer the range 2-8.

We saw that the median tree-level is sensitive to the $\beta$ value. While we do not have a standard to evaluate which levels are better for users, we know that the more a user prefers recall, the closer to the root, the less effort the user needs to make to reach that level.

We also saw that the Boolean query and CCIR retrieve different documents (Figure 3.3), and these documents could be relevant (analysis of SR80). Therefore, CCIR could serve users willing to use more than one IR method by finding more relevant documents.

### 3.5.2 What types of SRs are best served by CCIR?

We saw that there is a substantial variance among the F-score difference values of the SRs (Figure 3.2E), meaning that for some SRs, CCIR performs much worse than for others. We would imagine that, for CCIR, a SR with general needs (e.g. a disease) would perform better than a SR with specific needs (e.g. interaction between two medications), while the opposite would be true for Boolean queries (Carmel et al. [33] analyzed how the needs affect query difficulty). However, the three SRs that we analyzed had specific needs (Table 3.2) yet one had bad performance and two had good performance. The only clue that we can use to infer the performance of a SR is in SR47: its need is hard to write as a Boolean query, so we can infer that IR methods not based on a Boolean query are likely to have an advantage. However, this inference is more about the bad performance of the Boolean query than the good performance of the CCIR.

### 3.5.3 What are the strengths and weaknesses of CCIR?

#### 3.5.3.1 Strengths

CCIR may find documents that the Boolean query does not. We know this from the results of intersection proportions (Figure 3.3), where it shows that CCIR and the Boolean query retrieve different documents. We also know this from the newly discovered relevant document of SR80.

CCIR may reduce the noise of searches that are hard to write as a Boolean query. We know this from how CCIR performs well for SR47 and SR80: The former's Boolean query could not be sufficiently specific because the Boolean query format does not allow to specify subject-object relations between terms. The latter's Boolean query could not be specific because of the risk of missing documents with poorly annotated metadata.

CCIR may help expand the vocabulary used in a Boolean query. We know this from our experience with SR80. By looking at the difference between the noun-phrases of the parent cluster and the Boolean query of SR80, we realized that the Boolean query was missing a relevant search term which was likely not considered by the authors of the Boolean query.

#### 3.5.3.2 Weaknesses

CCIR may not create a cluster with the exact topic that the user needs. We know this because in SR47 and SR59 there was a divergence between the user needs and the topic of the CCIR sets of retrieved documents. The tree hierarchy did not had a cluster with the same topic as the user needs, which may happen because documents may relate to multiple topics.

The performance for a given SR can be unpredictable. We know this because of the high dispersion of the F-score difference values (Figure 3.2E) and because the characteristics of SR59, SR47 and SR80 did not give a clue about their performance.

Documents that share the same topic may be poorly directly or indirectly connected in a citation network. We know this from our experience with SR80. While a cluster with the relevant topic was retrieved, several relevant documents were missing. Also, the noun-phrases differences between the retrieved documents of the optimal cluster and the parent cluster of SR47 suggest that the optimal cluster was created based on the citation practices of the authors instead of the topic of the documents. Potentially, this issue could be diminished by combining citation-based and semantic-based clustering.

The clusters at the highest levels have too many documents, which makes the topic of the clusters hard to interpret for a real user because the documents are so diverse. This is a serious problem because selecting the wrong cluster at this level is a critical mistake [171]. Our evaluation did not suffer from this issue because CCIR already knows in which clusters the relevant documents can be found. In a real situation, a user may be able to handle this issue if they know at least some of the relevant documents, and then they could even select clusters bottom-up instead of top-down [161]. Alternatively, the user can create the tree hierarchy with fewer documents.

### 3.5.4 Limitations of this work

We identified four potential limitations to our work.

First, we did not cover all the possible clustering solutions. We used a single clustering solution, instead of using several clustering solutions or letting a user create clustering solutions on the run. Some of the characteristics of the tree hierarchy could have been different, like the clustering algorithm that we used, the clustering resolution parameters, the number of child clusters, the number of levels and the fact that we created the tree hierarchy by a top-down division of clusters instead of a bottom-up agglomeration of clusters.

Second, we did not cover all the possible citation networks. We used a citation network of direct citations, and not a more densely connected citation network using co-citations [145] or bibliographic coupling [96], which when combined with direct citation improve the representation of the structure of science [165]. We made the citation network using the full corpus, but we could also have used for example the documents retrieved from a query, which some studies reported to be more effective for cluster-based IR [152].

Third, the cluster selection algorithm does not reflect fully realistic (and noisy) user behavior. The cluster selection algorithm knows the relevant documents – an assumption commonly made in information retrieval evaluation –, which a real user would not. A real user would have to select the clusters based on their own personal evaluation of which cluster is more likely to contain the relevant documents, and also they would have to evaluate when to stop going down the tree hierarchy. This process would take cognitive effort, which our evaluation does not consider. A less cognitively heavy alternative for a user could be to eliminate a cluster that does not contain relevant documents and then create a new clustering solution, as there is likely to be an obvious candidate for elimination. This is the same process as selecting more than one cluster, as we discussed in the weaknesses (Section 3.5.3.2), and we decided against implementing it in the evaluation because it would create too many steps and the clustering would take too much computational resources. Another unrealistic behavior is that the cluster selection algorithm never chooses the wrong cluster, unlike a real user. We could have implemented mistakes by giving imperfect information to the cluster selection algorithm, but we decided not to so to have less variables that could affect the interpretation of our results. Finally, it is not realistic to allow the cluster selection algorithm to choose very small clusters (size between 1 and 10 documents) because this size of clusters does not appear in real situations (as discussed by Willet [171]). Future work could address more noisy user behavior, similar to user behavior modeling in information retrieval [81].

The final limitation is that it could be argued that the Boolean queries we used are not realistic. A real Boolean query is created over several iterations, where the creators of the query keep refining the query until they are satisfied with the search results. Our evaluation does not consider this. Also, our Boolean queries had a recall of 1.0 (i.e., they found all the relevant documents), which is unlikely for a real IR method. Additionally, we only considered the documents retrieved by the Boolean query on MEDLINE, while the authors of the SRs usually used more than one database or method to search for documents, including the expert knowledge of their colleagues. We did not include more sources because it would be too much effort to retrieve the documents of each method and to harmonize the results between SRs that used different methods. Finally, the translation from OVID format to PubMed format is likely to have modified the set of retrieved documents, especially if the Boolean query used OVID-specific features (like distance between words). We tried to remove the cases with the biggest modification of the set of retrieved documents by removing the SRs with Boolean queries that retrieved a number of documents too different from the number documents self-reported by the authors (see Section 3.3.1).

## 3.6 Conclusion

In this work we have shown some of the advantages and limitations of using CCIR for academic search, both for generic CCIR and for our specific tree hierarchy implementation. We have also introduced an evaluation protocol for cluster-based IR methods with the task of finding relevant documents for SRs. This protocol can be used and modified by other researchers. We release our data for use by other researchers in the form of the three tree hierarchies, the set of relevant documents and the set of documents retrieved by the Boolean query, the latter one created through intensive manual annotation. The current CCIR implementation can be used as a straightforward CCIR tool of value for real users.

Our research shows that the best served users are those who prefer recall over precision 2 to 8 times. Users that prefer even more recall, like SR users, are less well served, and users that prefer more precision are the worst served. CCIR may complement Boolean query searches in various ways: it may help SR users that have problems to state their requirements as Boolean queries, it may suggest terms for Boolean queries, and it may retrieve relevant documents not retrieved by a Boolean query.

A problematic aspect of CCIR is that performance varies significantly because there sometimes is no cluster that contains the topic of the SR. This may happen because documents may relate to multiple topics, leading to clusters that do not match with the topic of the SR. It may also happen because of a lack of citation connections between the documents related to the topic of interest. Another problematic aspect is that the current implementation of CCIR demands a high cognitive effort from a user.

For future work related to CCIR, interesting research directions are how to improve its performance (how to create better clusters, re-clustering based on the selection of multiple clusters by a user, mixing with semantic-based clustering), how it compares to other IR methods (especially citation-based or cluster-based methods) and how real users interact with it (how to select clusters, how to complement with other IR tools).

## 3.7 Data availability

The code used to run the experiments in this paper is available in GitHub (`https://github.com/jpbascur/citation_clusters_evaluation`) and the data and supplementary material is available in Zenodo [14].

## 3.8 CRediT author statement

**Juan Pablo Bascur:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.
**Suzan Verberne:** Conceptualization, Methodology, Supervision, Writing – review & editing.
**Nees Jan van Eck:** Conceptualization, Methodology, Software, Supervision, Writing – review & editing.
**Ludo Waltman:** Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

## 3.9 Acknowledgements

We would like to thank Jan W. Schoones for his expert support in biomedical Boolean queries, and Vincent Traag and Roel van der Ploeg for their invaluable feedback. We are also grateful to an anonymous reviewer for their comments on our work.

# Chapter 4

# Which topics are best represented by science maps? An analysis of clustering effectiveness for citation and text similarity networks

## Abstract[1]

A science map of topics is a visualization that shows topics identified algorithmically based on the bibliographic metadata of scientific publications. In practice not all topics are well represented in a science map. We analyzed how effectively different topics are represented in science maps created by clustering biomedical publications. To achieve this, we investigated which topic categories, obtained from MeSH terms, are better represented in science maps based on citation or text similarity networks. To evaluate the clustering effectiveness of topics, we determined the extent to which documents belonging to the same topic are grouped together in the same cluster. We found that the best and worst represented topic categories are the same for citation and text similarity networks. The best represented topic categories are diseases, psychology, anatomy, organisms and the techniques and equipment used for diagnostics and therapy, while the worst represented topic categories are natural science fields, geographical entities, information sciences and health care and occupations. Furthermore, for the diseases and organisms topic categories and for science maps with smaller clusters, we found that topics tend to be better represented in citation similarity networks than in text similarity networks.

## 4.1 Introduction

Science maps [38] are visualizations that provide an overview of the content of collections of scientific publications. The goal of science mapping is to find meaningful structures in the bibliographic metadata of publications (e.g, in the references, the titles and abstracts, or the authors). These structures can then be used for literature analysis or information retrieval [42, 154]. Some of the uses of science maps are field delimitation [184], research policy [149], and enhanced document browsing [17]. A well established practice to create science maps is to cluster similar publications, and then to summarize the content of the resulting clusters. Our focus in this paper is on science

[1] This chapter is based on: Juan Pablo Bascur, Suzan Verberne, Nees Jan van Eck and Ludo Waltman. 2025. Which topics are best represented by science maps? An analysis of clustering effectiveness for citation and text similarity networks. Scientometrics 130, 1181–1199. https://doi.org/10.1007/s11192-024-05218-6. [19]

maps created in this way.

When using science maps, it is important to be aware that scientific publications usually have more than a single topic (e.g., a document about the topic *lung cancer* is, implicitly, also about both *lungs* and *cancer*), but in a science map they typically can be assigned to only one cluster, where the cluster is intended to represent a single cohesive topic. Because in reality, publications can have more than one topic, losing information when creating science maps is unavoidable, but it does raise the question of which of the topics addressed in a collection of publications a clustering will be based on. This is not an idle question, as there can be significant disagreement between expert-identified and cluster-identified topics [78], indicating that expert-identified topics are poorly represented by the clusters in a science map. More specifically, an expert with an interest in a particular topic may find that publications related to this topic are scattered over many different clusters, with most of the publications in these clusters being unrelated to the expert's topic of interest. By providing a better understanding of the types of topics that are well or less well represented in science maps, we hope our research will contribute to a more effective use of these maps.

In this paper, we use the Medical Subject Headings (MeSH) terms to investigate clustering for biomedical topics. Our focus is on clustering solutions based on either citation or text similarity networks, which are the most common document similarity metrics for creating science maps. We aim to find out which MeSH terms are well represented by the clusters in a science map, a phenomenon that we will refer to as *clustering effectiveness*. Our approach is to group topics, represented by MeSH terms, into topic categories, represented by branches of the MeSH tree, and to then evaluate clustering effectiveness at the level of these topic categories.

Our research questions are as follows:

- Which topic categories have the highest and lowest clustering effectiveness in citation and text similarity networks?
- Which topic categories have higher clustering effectiveness in citation similarity networks than in text similarity networks, and vice versa?

In the remainder of this paper, we will discuss background literature, describe our data, define our metrics, report our analyses and discuss our results.

## 4.2 Background

This section has the following structure: In Section 4.2.1 we explain how science maps are usually evaluated, in Section 4.2.2 we explore the criticism of science maps that originates from one particular evaluation method, and in Section 4.2.3 we explain the challenges of understanding the meaning of the clusters in a science map.

### 4.2.1 Evaluation of science maps

In the current paper we evaluate the quality of science map only from the perspective of its field delimitation function. However, it is important to keep in mind that science maps are richer tools, with various features that can be interpreted beyond the extend to which clusters correspond to topics. For example, it can be evaluated on the extend to which the labels of the clusters and the distance between clusters provide useful visual information, or on how cross-cluster topics inform on the structure of the topics. The most common method to evaluate the quality of the field delimitation function a science map is to ask experts if the science map reflects their knowledge of the field of interest. The utility of this evaluation method has recently been called into question because it usually gives an inconclusive result: The experts tend to agree with most of the science map but identify caveats about certain details [64]. Additionally, there are several issues intrinsic to the expert evaluation method: The evaluation criteria may differ between experts; seeing the map may affect the expert's understanding of a field; the expert may be biased towards the subfields of their interest; and the expert may have limited competence in some subfields [64].

An alternative method to evaluate the quality of a science map is to consider the intrinsic properties of the clustering process used to create science maps. Commonly used intrinsic properties are desirable characteristics such as homogeneous cluster sizes, few small clusters, stable clustering solutions between different runs of the cluster algorithm, and a short computing time to create the clusters [148]. An intrinsic properties evaluation method was developed by Waltman et al. [165]. Their method assumes that there exists an ideal map and then assesses how closely a clustering solution matches this map. It evaluates the quality of a clustering solution based on one metric using another unrelated baseline metric (e.g., a clustering solution based on citation similarity can be evaluated using text similarity). Ahlgren et al. [4], who created the clustering solutions that we use in our current work, used this method with MeSH terms similarity as their baseline metric.

A third approach to evaluate the quality of a science map is to define a ground truth made of documents that correspond to a given topic, and evaluate the overlap between the clustering solution and the ground truth: either the extent to which all documents of each field are contained in a single cluster [77, 78], or the extent to which each cluster contains only documents of a single field [69, 76, 78, 135]. Some studies obtained the ground truth from the references of review articles [98, 142], but most studies obtained the ground truth using expert knowledge. To our knowledge, MeSH terms have not been used as ground truths, although Sjögårde, Ahlgren and Waltman [144] used MeSH terms to label clusters in science maps. It is worth mentioning that our work has a different goal than evaluating a science map based on a ground truth. Instead of evaluating the quality of a science map based on a set of topics, we evaluate which topic categories are most accurately represented in a science map.

### 4.2.2 Criticism of science maps based on ground truth evaluations

Evaluations that use expert knowledge ground truths have recently questioned the quality of science maps by challenging their ability to identify fields of science [69, 76–78]. For example, Held and Velden [76] found that science maps provide clusters about organisms rather than clusters about the field of invasive biology. One explanation for these negative results is that a document can belong to several fields or topics but only to a single cluster [76, 78] (although some maps allow documents to belong to multiple clusters [70, 178]). Another explanation is that the choice of a clustering algorithm can have a significant influence on the quality of a science map, and it is impossible to know beforehand which clustering algorithm will give the best result for a given map [74, 135].

Similar negative findings have also emerged in areas beyond science mapping. For example, the field of complex systems has developed algorithms to clusters the elements that share a given property (i.e., the cluster matches the ground truth), but these algorithms fail in practical applications. On the other hand, this field has succeeded in practical applications of algorithms that infer the properties of an element based on the properties of the other elements in a cluster (e.g., fraud in telecommunications networks, function in biological networks) [62, 86, 125].

### 4.2.3 Meaning of the clusters

The negative findings discussed in the previous section suggest that science maps, and clustering in general, offer poor representations of certain ground truths. However, this does not mean that science maps are not useful. As mentioned in Section 4.2.1, experts tend to agree that science maps reflect their knowledge of a field. Also, in the field of complex systems, Newman and Clauset [121] argued that, even if clusters do not reflect the ground truth, they can still describe meaningful structures in the data. Our work tries to find out what kinds of structures are described by the clusters in a science map.

In this direction, Seitz et al. [141] found that the epistemic functions of citations (i.e., what kind of knowledge is a citation contributing to in a document) within a cluster are different from the epistemic functions of citations between clusters. This suggests that clusters tend to represent certain epistemic functions more than others. Also, the type of similarity network might have an effect on the meaning of clusters. For example, Ding [52] found significant differences between clusters emerging from co-authorship networks of documents and clusters emerging from topic modeling of documents. On the other hand, Velden et al. [162] found that there is a substantial similarity between the topics found in science maps built from citation and text similarity networks, although science maps built from citation networks are better at distinguishing topics when words related to the topics have multiple meanings.

## 4.3 Methods

This section has the following structure: In Section 4.3.1, we define how we selected our data. In Section 4.3.2, we explain how we modified our data so to better fit our experimental design. In Section 4.3.3, we explain how we evaluate the clustering effectiveness of topic categories.

### 4.3.1 Data selection

**Documents** The collection of documents that we use in our work comes from the work by Ahlgren et al. [4]. This is a collection of 2,941,119 PubMed documents published between 2013 and 2017.

**Clustering solutions** The clustering solutions that we use are the ones generated by Ahlgren et al. They created several clustering solutions for the above mentioned documents using different similarity metrics and granularities. They used the Leiden algorithm [153] for clustering, where the parameter Resolution controls the granularity of the clustering solution (a higher Resolution value generates smaller clusters). We select two similarity metrics, one for citation and one for text, based on which pair of metrics produce similar cluster sizes at the same Resolution. The citation metric is *Extended direct citation*, which is calculated using direct citations between documents plus the citations to documents outside the document collection [165]. The text metric is *BM25* [133], which uses the noun phrases in the titles and abstracts of the documents, and weights them inversely to their frequency in the document collection [165]. For each metric we selected the three clustering solutions that use the Resolution values $2*10^{-6}$, $2*10^{-5}$ or $2*10^{-4}$, enabling us to evaluate different cluster sizes. We selected these Resolution values because the first and second value yield cluster sizes similar to those in the algorithmic mapping of science [164] used in the CWTS Leiden Ranking [35], while the third value enables us to evaluate clusters of smaller size.

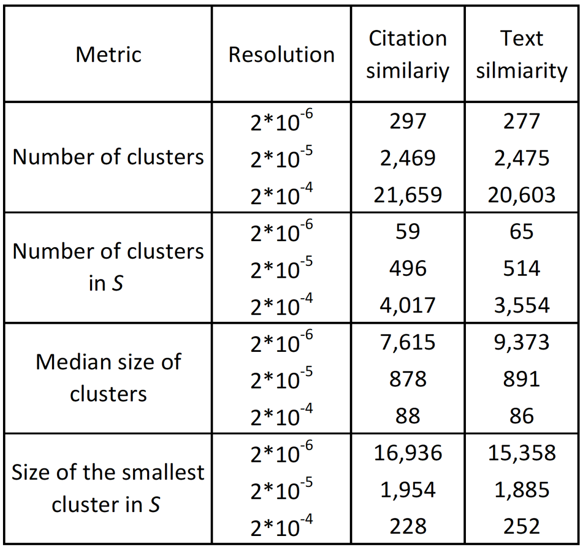
**Topics** Our topics are the MeSH terms, a controlled vocabulary thesaurus from the National Library of Medicine (NLM) used for indexing PubMed. MeSH terms are semi-automatically annotated to documents by the NLM [117]. We obtained the MeSH terms annotated for each document in our document collection, plus the metadata of the MeSH terms themselves, from the PubMed and MeSH databases (version from 2023) available in the database system of the Centre for Science and Technology Studies (CWTS) at Leiden University.

**Topic categories** Our topic categories are the 16 nodes at the first level of the MeSH hierarchical tree of topics [117], also known as the branches of the MeSH tree. We use branches because they group the MeSH terms in epistemological categories (e.g., organisms), which are the categories sometimes used for topical analysis of clusters [78, 141]. A single MeSH term can have instances in different branches of the MeSH tree. We will address this in Section 4.3.2.

### 4.3.2 Data prepossessing

**Clustering solution cleaning** We cleaned the clustering solutions by removing the clusters with fewer than 10 documents because these clusters usually had documents that were disconnected from the largest connected component of the similarity network. Removing these clusters removed only a minor fraction of the total number of documents. The statistics of each clustering solution after this process can be seen in Table 4.1. In this table, the variable $S$ is the smallest set of clusters

Table 4.1: Statistics of the clustering solutions. $S$ is the smallest set of clusters that together cover at least half of the documents in the dataset. The size of the cluster is the number of documents it contains.



| Metric | Resolution | Citation similariy | Text silmiarity |
|---|---|---|---|
| Number of clusters | $2*10^{-6}$ | 297 | 277 |
| | $2*10^{-5}$ | 2,469 | 2,475 |
| | $2*10^{-4}$ | 21,659 | 20,603 |
| Number of clusters in *S* | $2*10^{-6}$ | 59 | 65 |
| | $2*10^{-5}$ | 496 | 514 |
| | $2*10^{-4}$ | 4,017 | 3,554 |
| Median size of clusters | $2*10^{-6}$ | 7,615 | 9,373 |
| | $2*10^{-5}$ | 878 | 891 |
| | $2*10^{-4}$ | 88 | 86 |
| Size of the smallest cluster in *S* | $2*10^{-6}$ | 16,936 | 15,358 |
| | $2*10^{-5}$ | 1,954 | 1,885 |
| | $2*10^{-4}$ | 228 | 252 |

that together cover at least half of the documents in the dataset. This means that $S$ contains the biggest clusters in the clustering solution. We report statistics for $S$ to provide some insight into the distribution of cluster sizes.

**MeSH term expansion** We would like a MeSH term to be annotated on all documents related to the topic of the MeSH term, but NLM typically only annotates up to 15 MeSH terms per document, which means that more generic MeSH terms are not annotated. To fix this, we expanded the number of MeSH terms annotated to a document by annotating, for each NLM MeSH term, all MeSH terms that are upstream in the MeSH tree, or in other words, all ancestors of the NLM MeSH term in the MeSH tree.

For example, if a document has the NLM MeSH term *Abdominal Pain*, we also annotated the upstream MeSH term *Pain*. While the former MeSH term belongs to the branch Diseases [C], the latter one belongs not only to the branch Diseases [C], but also to the branches Psychiatry and Psychology [F] and Phenomena and Processes [G]. We annotated the MeSH term *Pain* paired with the branch Diseases [C], and not with the other two branches. On the other hand, if a document has the NLM MeSH term *Pain*, then we would annotate three versions of it, one for each branch. For simplicity, in the rest of this paper we will refer to MeSH terms paired with a specific branch simply as MeSH terms. Also, we will refer to the documents that have a given MeSH term as the MeSH term documents and to the number of these documents as the MeSH term size.

**MeSH term removal** We removed some MeSH terms to improve the quality of our experiments. Our first removal criterion is size. We removed MeSH terms with size greater than 300,000 (i.e., 10% of the document set) because these MeSH term documents can saturate the clusters just by random chance, distorting our analysis. We also removed the MeSH terms with size 500 or less, because we want the smallest MeSH terms to be close but smaller than the median size of the clusters for resolution $2 * 10^{-5}$.

Our second removal criterion is redundancy. Due to the MeSH term expansion process, some MeSH terms had almost the same documents as their ancestor in the MeSH tree, like *Dogs* and its ancestor *Canidae*. This redundancy could distort our results. We therefore decided to remove the redundant MeSH terms by grouping together MeSH terms that share many documents and retaining

Table 4.2: Number of MeSH terms per branch and Size bin. A Size bin is a range of topic sizes. A topic size is the number of documents in the topic.

| Branch | Size bin | | | | | |
|---|---|---|---|---|---|---|
| | 501 - 1,000 | 1,001 - 2,000 | 2,001 - 4,000 | 4,001 - 8,000 | 8,001 - 16,000 | Total |
| Anatomy [A] | 209 | 201 | 161 | 102 | 76 | 749 |
| Organisms [B] | 247 | 168 | 98 | 75 | 44 | 632 |
| Diseases [C] | 472 | 391 | 272 | 194 | 114 | 1,443 |
| Chemicals and Drugs [D] | 1,033 | 785 | 568 | 357 | 264 | 3,007 |
| Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E] | 324 | 298 | 253 | 189 | 150 | 1,214 |
| Psychiatry and Psychology [F] | 109 | 113 | 95 | 65 | 38 | 420 |
| Phenomena and Processes [G] | 264 | 244 | 221 | 179 | 143 | 1,051 |
| Disciplines and Occupations [H] | 50 | 28 | 31 | 23 | 15 | 147 |
| Anthropology, Education, Sociology, and Social Phenomena [I] | 57 | 56 | 40 | 29 | 24 | 206 |
| Technology, Industry, and Agriculture [J] | 76 | 70 | 68 | 24 | 26 | 264 |
| Information Science [L] | 31 | 35 | 28 | 20 | 18 | 132 |
| Named Groups [M] | 21 | 34 | 20 | 11 | 14 | 100 |
| Health Care [N] | 182 | 150 | 134 | 110 | 87 | 663 |
| Geographicals [Z] | 51 | 39 | 36 | 16 | 21 | 163 |
| Total | 3,126 | 2,612 | 2,025 | 1,394 | 1,034 | 10,191 |

only the smallest MeSH term from the group, which in our experience tends to be the term that best represents the group. The extent to which MeSH terms share documents was measured using Jaccard similarity, the grouping algorithm was agglomerate hierarchical clustering with the Complete Linkage method [137], and the criterion for forming MeSH term groups was for MeSH terms to have a Jaccard similarity of at least 0.9. In cases where a group had more than one smallest MeSH term, we selected the one at the lowest level in the MeSH tree or the one with the largest number of instances in the MeSH tree.

**Branch removal** To make our results more robust, we removed the branches with fewer than 100 MeSH terms. We ended up with the 14 branches shown in Table 4.2.

**Size bins of MeSH terms** The size of a MeSH term can be expected to have an effect on its clustering effectiveness. We therefore grouped the MeSH terms according to their size. We refer to these groups as Size bins. To ensure the robustness of our results, we only considered Size bins that had at least 10 MeSH terms per branch. This resulted in five Size bins: 501-1,000, 1,001-2,000, 2,001-4,000, 4,001-8,000, and 8,001-16,000. The number of MeSH terms per Size bin can be seen in Table 4.2.

### 4.3.3 Clustering effectiveness

**Selection of clusters** To find out which MeSH terms are well represented by the clusters in a science map, we introduce the notion of clustering effectiveness. Measuring the clustering effectiveness of a MeSH term starts by selecting a subset of clusters. Our cluster selection criterion is to select the clusters with the largest number of MeSH term documents while making sure that the selected clusters cover at least a given share of all MeSH term documents. We call this share Coverage. We consider three Coverage values: 0.25, 0.50 and 0.75. Our cluster selection criterion minimizes the number of selected clusters for a given Coverage value. It is inspired by cluster quality metrics of Yuan, Zobel and Ling [181]. We expect our cluster selection criterion to reflect the clusters a user of a science map is likely to select while exploring the map.

**Clustering effectiveness metrics** Once we have the selected clusters for a given MeSH term, we measure clustering effectiveness using two metrics:

- Purity: Purity represents the extent to which the selected clusters are composed of MeSH term documents. It is the fraction of documents in the selected clusters that are MeSH term documents. In mathematical terms, Purity is defined as:

  $$Purity = \frac{\sum_{i=1}^{N} |D_i \cap D_M|}{\sum_{i=1}^{N} |D_i|} \tag{4.1}$$

  Here, $N$ denotes the number of selected clusters, $D_i$ denotes the documents in selected cluster $i$ and $D_M$ denotes the MeSH term documents. The higher Purity, the more effective the clustering. Purity is bounded between zero and one.

- Inverse count of clusters (ICC): ICC represents the extent to which the MeSH term documents are contained in a small number of clusters. ICC is defined as one divided by the number of selected clusters. In mathematical terms, ICC is defined as:

  $$ICC = \frac{1}{N} \tag{4.2}$$

  The higher ICC, the more effective the clustering. Like Purity, ICC is bounded between zero and one.

We use two metrics instead of one to control for MeSH term size and cluster size: If there are few MeSH term documents, or if they are in big clusters, then ICC will be high but Purity will be low, and vice versa.

The Purity and ICC of a MeSH term are calculated for a given Coverage value, Resolution value and similarity network. We use C-Purity and C-ICC to refer to Purity and ICC calculated for a citation network, and T-Purity and T-ICC to refer to Purity and ICC calculated for a text network.

We also provide metrics for the difference in Purity and ICC between citation and text networks for a given MeSH term. These metrics, referred to as rPurity (Ratio Purity) and rICC (Ratio ICC), are calculated as the logarithm base 2 of C-Purity or C-ICC divided by T-Purity or T-ICC. The purpose of the logarithm is to facilitate the interpretation of the results (e.g. for rPurity vale -1, T-Purity is double C-Purity, and for +1 is the opposite). In mathematical terms, rPurity and rICC are defined as:

$$rPurity = \log_2 \left( \frac{C\text{-}Purity}{T\text{-}Purity} \right) \tag{4.3}$$

$$rICC = \log_2 \left( \frac{C\text{-}ICC}{T\text{-}ICC} \right) \tag{4.4}$$

Positive values indicate that a citation network yields a higher clustering effectiveness than a text network, and vice versa.

## 4.4 Results

### 4.4.1 Which topic categories have the highest and lowest clustering effectiveness in citation and text similarity networks?

To answer our first research question, we consider the C-Purity and T-Purity rankings of the 14 branches for each of the 45 combinations of parameter values (i.e., three Resolution values combined with three Coverage values combined with five Size bin values). Table 4.3 shows the number of times each branch appears in each position in the C-Purity and T-Purity rankings. The order of the branches in the table was determined manually so that the branches that frequently occupy higher ranking positions are above of the ones that occupy lower ranking positions. We found that the ICC rankings are strongly correlated with the Purity rankings, so we do not show them.

From Table 4.3 we make the following observations:

- Most of the branches occupy between one and four adjacent positions, which shows that the position of the branches tends to be stable for different parameter values.
- For both C-Purity and T-Purity, the top five branches are almost always in positions 1 to 7, and the bottom four branches are almost always in positions 8 to 14. We therefore consider the top five and bottom four branches as the the ones with, respectively, the highest and lowest clustering effectiveness.
- The top five and bottom four branches are the same for C-Purity and T-Purity, showing that in this respect citation and text networks yield very similar outcomes.
- The top five branches are Diseases [C], Organisms [B], Anatomy [A], Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E] and Psychiatry and Psychology [F].
- The bottom four branches are Health Care [N], Disciplines and Occupations [H], Information Science [L] and Geographicals [Z].

Figure 4.1 shows the distribution of the Purity and ICC values of each branch for the 45 combinations of parameter values. The box plots for the different branches heavily overlap with each other due to the effect of the parameter values on Purity and ICC. From Figure 4.1 we observe that C-Purity, T-Purity, C-ICC and T-ICC are substantially higher for the branch Diseases [C] than for the other branches, while they are substantially lower for the branch Geographicals [Z]. This also explains why in Table 4.3 these branches almost always appear in position 1 and 14, respectively.

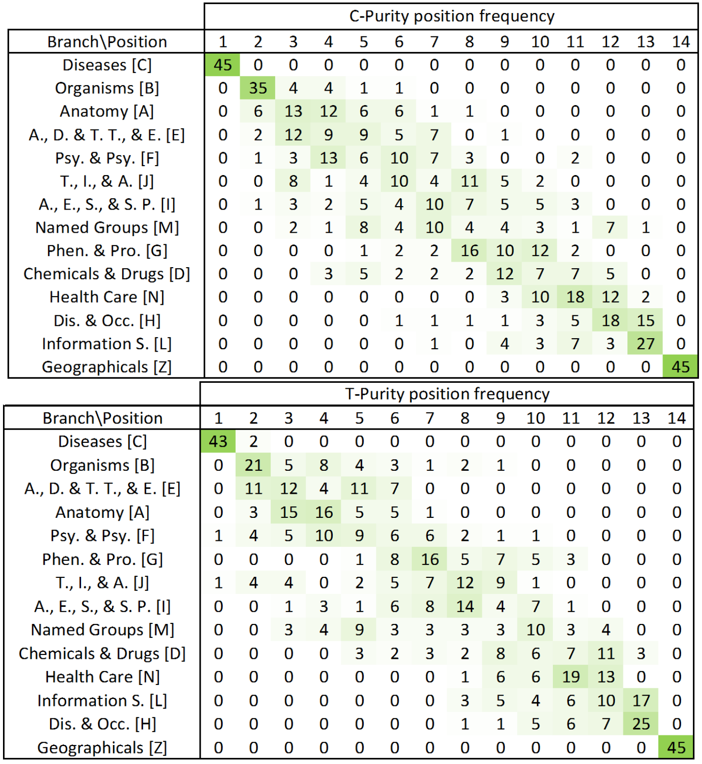
### 4.4.2 Which topic categories have higher clustering effectiveness in citation similarity networks than in text similarity networks, and vice versa?

To address our second research question, we first evaluate how the ratio metrics rPurity and rICC correlate with the Size bin, Resolution and Coverage parameters. The box plots in Figure 4.2 show the distribution of the rPurity and rICC values for each value of the Size bin, Resolution and Coverage parameters. Here we see that higher Resolution and Coverage are correlated with higher rPurity and rICC. Also, higher Size bin is correlated with lower rPurity and rICC, but this is a weak correlation.

The answer to our second research question depends on whether the rPurity and rICC values of a branch are positive or negative. Positive values indicate that the clustering effectiveness is higher in citation networks, while negative values indicate that the clustering effectiveness is higher in text networks. The box plots in Figure 4.3 show the distribution of the rPurity and rICC values

Table 4.3: .

Number of times each branch appears in each ranking position, using either C-Purity (top) or T-Purity (bottom) as ranking criterion.



| | C-Purity position frequency | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Branch\Position | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| Diseases [C] | 45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Organisms [B] | 0 | 35 | 4 | 4 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Anatomy [A] | 0 | 6 | 13 | 12 | 6 | 6 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| A., D. & T. T., & E. [E] | 0 | 2 | 12 | 9 | 9 | 5 | 7 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| Psy. & Psy. [F] | 0 | 1 | 3 | 13 | 6 | 10 | 7 | 3 | 0 | 0 | 2 | 0 | 0 | 0 |
| T., I., & A. [J] | 0 | 0 | 8 | 1 | 4 | 10 | 4 | 11 | 5 | 2 | 0 | 0 | 0 | 0 |
| A., E., S., & S. P. [I] | 0 | 1 | 3 | 2 | 5 | 4 | 10 | 7 | 5 | 5 | 3 | 0 | 0 | 0 |
| Named Groups [M] | 0 | 0 | 2 | 1 | 8 | 4 | 10 | 4 | 4 | 3 | 1 | 7 | 1 | 0 |
| Phen. & Pro. [G] | 0 | 0 | 0 | 0 | 1 | 2 | 2 | 16 | 10 | 12 | 2 | 0 | 0 | 0 |
| Chemicals & Drugs [D] | 0 | 0 | 0 | 3 | 5 | 2 | 2 | 2 | 12 | 7 | 7 | 5 | 0 | 0 |
| Health Care [N] | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 10 | 18 | 12 | 2 | 0 |
| Dis. & Occ. [H] | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 3 | 5 | 18 | 15 | 0 |
| Information S. [L] | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 4 | 3 | 7 | 3 | 27 | 0 |
| Geographicals [Z] | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 45 |

| | T-Purity position frequency | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Branch\Position | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| Diseases [C] | 43 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Organisms [B] | 0 | 21 | 5 | 8 | 4 | 3 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 |
| A., D. & T. T., & E. [E] | 0 | 11 | 12 | 4 | 11 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Anatomy [A] | 0 | 3 | 15 | 16 | 5 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Psy. & Psy. [F] | 1 | 4 | 5 | 10 | 9 | 6 | 6 | 2 | 1 | 1 | 0 | 0 | 0 | 0 |
| Phen. & Pro. [G] | 0 | 0 | 0 | 0 | 1 | 8 | 16 | 5 | 7 | 5 | 3 | 0 | 0 | 0 |
| T., I., & A. [J] | 1 | 4 | 4 | 0 | 2 | 5 | 7 | 12 | 9 | 1 | 0 | 0 | 0 | 0 |
| A., E., S., & S. P. [I] | 0 | 0 | 1 | 3 | 1 | 6 | 8 | 14 | 4 | 7 | 1 | 0 | 0 | 0 |
| Named Groups [M] | 0 | 0 | 3 | 4 | 9 | 3 | 3 | 3 | 3 | 10 | 3 | 4 | 0 | 0 |
| Chemicals & Drugs [D] | 0 | 0 | 0 | 0 | 3 | 2 | 3 | 2 | 8 | 6 | 7 | 11 | 3 | 0 |
| Health Care [N] | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 6 | 6 | 19 | 13 | 0 | 0 |
| Information S. [L] | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 5 | 4 | 6 | 10 | 17 | 0 |
| Dis. & Occ. [H] | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 5 | 6 | 7 | 25 | 0 |
| Geographicals [Z] | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 45 |

of each branch for the 45 combinations of parameter values. For each branch, the rPurity and rICC distributions include both positive and negative values. This reflects the dependence of the rPurity and rICC values on the values of the Size bin, Resolution and Coverage parameters, as was shown in Figure 4.2.

Because for each branch the rPurity and rICC distributions include both positive and negative values, it is not possible to unequivocally conclude that a branch has a higher clustering effectiveness in either citation networks or text networks. Nevertheless, it is clear that the branches Diseases [C] and Organisms [B] tend to have a higher clustering effectiveness in citation networks than in text networks. rPurity and rICC are almost always positive for these branches. In contrast, the branches Geographicals [Z], Information Science [L], Named Groups [M], Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E] and Phenomena and Processes [G] tend to have a higher clustering effectiveness in text networks than in citation networks. However, the results for these branches are less stable, so we need to be cautious in drawing strong conclusions.

## 4.5 Discussion

This section has the following structure: We discuss what we have learned for our first research question in Section 4.5.1, for our second research question in Section 4.5.2, and for the strengths and weaknesses of our work in Section 4.5.3.

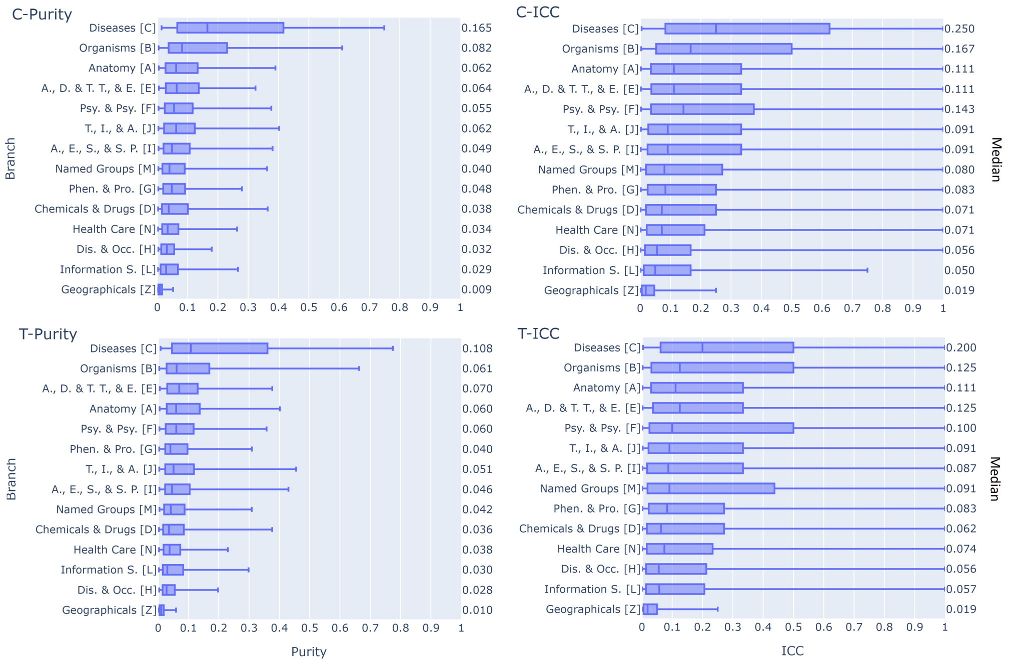


Figure 4.1: Box plots showing the distribution of C-Purity, C-ICC, T-Purity and T-ICC over the 45 combinations of parameter values. The median values of each box plot are reported along the right Y axis. The branches are sorted as in Table 4.3.

### 4.5.1 Which topic categories have the highest and lowest clustering effectiveness in citation and text similarity networks?

Our results show that the MeSH branches with the highest and lowest clustering effectiveness are the same for citation and text similarity networks. Despite the different purposes of writing and citing [104], the way scientists write and the way they cite yield similar rankings of MeSH branches in terms of clustering effectiveness. It would be interesting to see if the top and bottom branches are also the same in other similarity networks, like co-tweeting [45], co-authorship [120], and patent co-citation [102].

The branch Disciplines and Occupations [H], which contains the MeSH terms for natural science fields, is among the branches with the lowest clustering effectiveness. This suggests that how scientists cite each other is only weakly related to how they define scientific fields, which suggest the need for alternative approaches to defining scientific fields, for instance based on science map clusters. However, it is unclear to which extent this branch is a good representative of the natural science fields (e.g. the branch also includes MeSH terms about health occupations, and documents with NLM MeSH terms about natural science fields tend to be about meta-science). Therefore, a deeper analysis is required to support the suggestion, but this goes beyond the scope of the current paper.

Held and Velden [76] reported that a given science map was poor at showing the field of invasive biology, and instead placed documents related to the field in clusters about species. Our results are in line with this, because invasive biology belongs to Disciplines and Occupations [H], one of the bottom four branches in our results, while species belongs to Organisms [B], one of the top five branches.

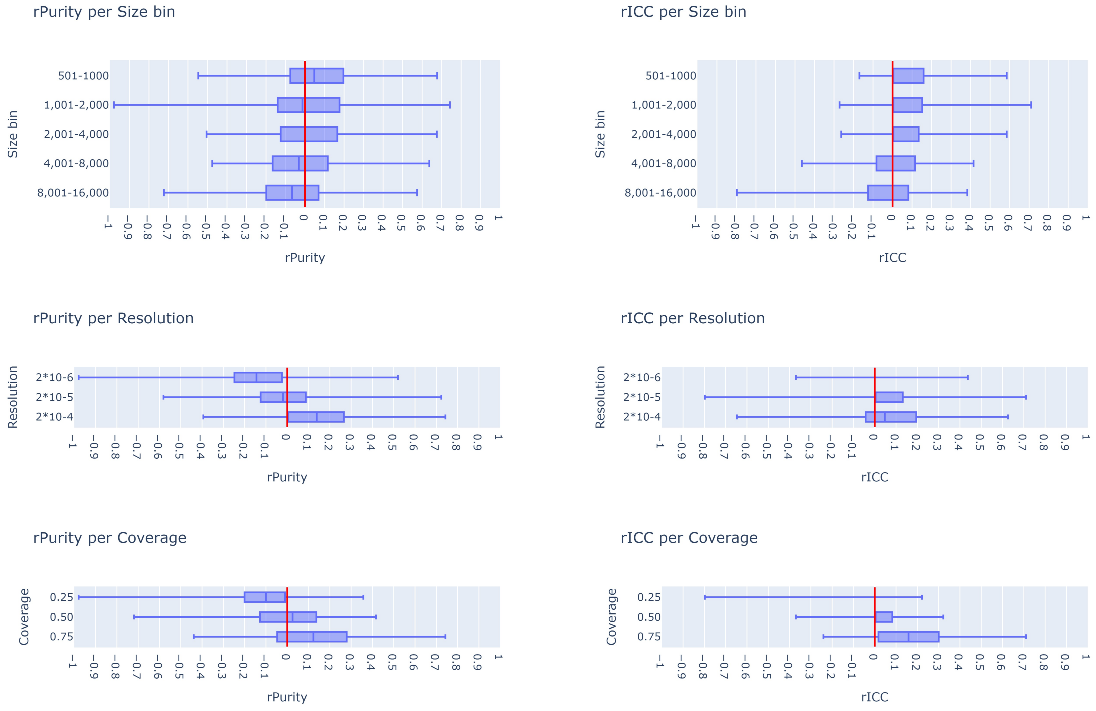


Figure 4.2: Box plots showing the distribution of rPurity and rICC for each value of Size bin, Resolution and Coverage.

### 4.5.2 Which topic categories have higher clustering effectiveness in citation similarity networks than in text similarity networks, and vice versa?

Our results show that which networks yield a higher clustering effectiveness depends strongly on the Resolution and Coverage values, with higher Resolution and higher Coverage increasing the clustering effectiveness for citation networks relative to text networks. Importantly, this does not mean that higher Resolution and higher Coverage increase the clustering effectiveness for citation networks in an absolute sense. It means that higher Resolution and higher Coverage increase the ratio between the clustering effectiveness for citation networks and the clustering effectiveness for text networks.

Ahlgren et al. [4] developed a method to measure the accuracy of the clusters in a science map. Using their data and visualization method, we found that the accuracy of citation networks relative to text networks increases as the Resolution value increases. This is in line with our results. Unfortunately, we do not know the mechanism behind this dependency. Our findings for Resolution could be useful for users of science maps: It tells them that, if they have two science maps, one based on citations and another based on text, then decreasing the size of the clusters will make the citation one more effective relative to the text one, and vice versa.

In the context of field delimitation tasks, where a user of a science map identifies the clusters that contain the documents of a field, Coverage is analog to the completeness of the field delimitation. Our findings for Coverage suggest that citation networks are better for exhaustive field delimitation, while text networks are better for less exhaustive field delimitation.

Our results also indicate that, omitting the effect of Resolution and Coverage, the branches Diseases [C] and Organisms [B] tend to have higher clustering effectiveness in citation networks than in text networks. To exemplify what this means for users, we consider the use case of Held and

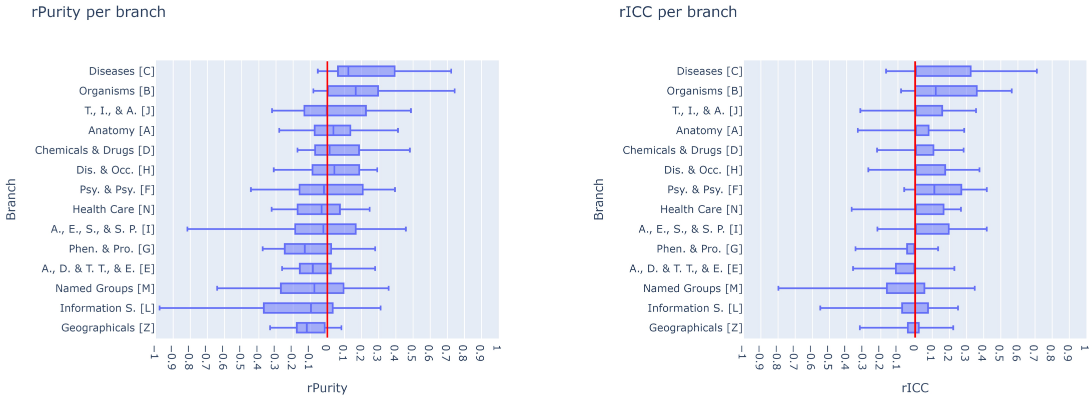


Figure 4.3: Box plots showing the distribution of rPurity and rICC for each branch.

Velden [76] discussed above: They would like to have a clustering of the field of invasive biology, but in their science map invasive biology documents are spread over clusters about organisms. If instead of a citation network a text network is used, the organisms will probably be clustered less effectively, which may give the opportunity for invasive biology documents to form their own clusters instead of being part of clusters about organisms.

### 4.5.3 Strengths and weaknesses

We see the use of MeSH terms as an important strength of our work. An alternative approach could be to ask experts to assign documents to topics, but this cannot be done at the scale at which MeSH terms provide document-topic links. Also, MeSH terms link documents to topics at a scale that no other classification scheme, like the Mathematics Subject Classification, the ACM Computing Classification System, or the Physics Subject Headings, is able to provide.

We also improved the utility of the MeSH terms by using Coverage, MeSH term expansion, MeSH term removal and MeSH branches in our experimental design. Coverage diminished the effect of mislabeled documents (e.g., the document with DOI *10.1007/s12603-020-1457-6* is incorrectly labeled with the MeSH term *Alcohol Drinking*) by ignoring a certain share of the documents with a particular MeSH term. MeSH term expansion allowed us to have a collection of documents for each MeSH term that represent the topic of the MeSH term more accurately. MeSH term removal allowed us to ensure that our results are not affected by redundant MeSH terms. Using the MeSH branches as topic categories allowed us to use a curated scheme of topic categories. However, some topic categories may be absent from the MeSH tree (e.g., topics linking diseases with their medicines) and some lower levels of the MeSH tree may be more informative as topic categories (e.g., the children of the branch Disciplines and Occupations [H] are *Natural Science Disciplines* and *Health Occupations*, which may be more informative as topic categories than the branch itself). It is worth mentioning that MeSH terms have an attribute (MeSH Major Topic) that indicates if the MeSH term is one of the major topics of the document. We did not use this attribute because only half of our documents had any MeSH term with this attribute.

Another strength of our work is that we evaluated clustering effectiveness per MeSH term, while other studies, like Waltman et al. [165], evaluated a clustering solution as a whole. Our method is also insensitive to the effect of size differences between MeSH terms and clusters (e.g., if clusters are much bigger than MeSH terms, it is impossible to have maximum Purity, and if they are much smaller, it is impossible to have maximum ICC) because our focus is on comparing the clustering effectiveness of different topic categories instead of achieving optimal clustering effectiveness.

A weakness of our work is that we used only one clustering algorithm, the Leiden algorithm, an algorithm that is commonly used by the science mapping community. Other studies used multiple algorithms: Held, Laudel and Gläser [77, 78] analyzed clusters created by the Leiden algorithm and the Infomap algorithm. Held [74] assessed the suitability of the Leiden, Louvian, OSLM and Infomap algorithms for creating clusters. Beyond science maps, Rossetti, Pappalardo and Rinzivillo [135] showed that different clustering algorithms (Louvain, Infohiermap, cFinder, Demon, iLCD and Ego-Network) have differential performance for different types of networks (DBLP co-authorship network, Amazon co-purchase network, YouTube users network, and LiveJournal users network).

Another weakness of our work is that we used only one citation similarity metric (extended direct citation) and only one text similarity metric (BM25). Future work should ideally evaluate multiple citation and text similarity metrics, because different citation metrics and different text metrics may yield different results.

A final weakness of our research is that our findings might be valid only for the current document set. Using document sets from other time periods or other fields (MeSH terms specialize in Biomedical fields) could have different results due to changes in the writing style and the epistemic functions of citations.

## 4.6 Conclusion

In this paper we explored science maps of mostly biomedical topics, analyzing the clustering effectiveness for different topic categories. We hope our work will contribute to a more effective use of science maps. We addressed the following research questions:

**Which topic categories have the highest and lowest clustering effectiveness in citation and text similarity networks?** We found that the answer is the same for citation and text similarity networks. Paraphrasing the topic category names, the topic categories with the highest clustering effectiveness are diseases, psychology, anatomy, organisms and the techniques and equipment used for diagnostics and therapy, while the topic categories with the lowest clustering effectiveness are natural science fields, geographical entities, information sciences and health care and occupations. Also, the diseases category has a substantially higher clustering effectiveness than all other categories, while the geographical entities category has a substantially lower clustering effectiveness.

**Which topic categories have higher clustering effectiveness in citation similarity networks than in text similarity networks, and vice versa?** We found that there are two factors that can make any topic category have higher clustering effectiveness in either network. The first factor is the size of the clusters generated by the clustering process (i.e., the Resolution parameter). The smaller the size, the higher the clustering effectiveness in citation networks relative to text networks. The second factor, specific to our experimental setting, is the percentage of all topic documents that must be covered by the selected clusters (i.e., the Coverage parameter). The higher this percentage, the higher the clustering effectiveness in citation networks relative to text networks. Regardless of these two factors, we found that the diseases and organisms topic categories tend to have higher clustering effectiveness in citation networks than in text networks.

Our work has shown that there is a strong tendency for clusters in science maps to represent some topics better than others. Further research could explore how to control which topics are clustered better, so that users of science maps can adjust the maps to their needs.

## 4.7 Data availability

The code used to run the experiments and the data needed to replicate the results are available in Zenodo [12].

## 4.8 CRediT author statement

**Juan Pablo Bascur:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.
**Suzan Verberne:** Conceptualization, Methodology, Supervision, Writing – review & editing.
**Nees Jan van Eck:** Conceptualization, Methodology, Supervision, Writing – review & editing.
**Ludo Waltman:** Conceptualization, Methodology, Resources, Supervision, Writing – review & editing.

# Chapter 5

# Use of diverse data sources to control which topics emerge in a science map

## Abstract[1]

Traditional science maps visualize topics by clustering documents, but they are inherently biased toward clustering certain topics over others. If these topics could be chosen, then the science maps could be tailored for different needs. In this paper, we explore the use of document networks from diverse data sources as a tool to control the topic clustering bias of a science map. We analyze this by evaluating the clustering effectiveness of several topic categories over two traditional and six non-traditional data sources. We found that the topics favored in each non-traditional data source are about: Health for Facebook users, biotechnology for patent families, government and social issues for policy documents, food for Twitter conversations, nursing for Twitter users, and geographical entities for document authors (the favoring in this latter source was particularly strong). Our results show that diverse data sources can be used to control topic bias, which opens up the possibility of creating science maps tailored for different needs.

## 5.1 Introduction

Science maps are a form of visualization that provides a content overview of a collection of academic documents. They are typically used for literature analysis [184], field delimitation, research policy, and enhanced document browsing [17]. A typical practice to create science maps is first to create a network of academic documents where the links are an aspect of the documents (e.g. bibliographic metadata), then to cluster together the documents that are well connected, and finally to summarize the contents of these clusters. In other words, the map is a set of clusters that emerge from document connections, and what a cluster represents is inferred from its documents.

In our previous work [19] we evaluated the extent to which a science map can place the documents of a topic inside clusters where most documents belong to that topic (i.e. to create clusters about the topic), a concept we refer to as clustering effectiveness. There, we found that the clustering effectiveness changes depending on the kind of topic, or in other words, that the maps have a bias toward clustering certain kinds of topics more effectively than others. For example, we found that in maps based on citation links or text similarity, topics related to diseases are well clustered while topics related to geographical locations are not. This bias can prove inconvenient for science map

[1] This chapter is based on: Juan Pablo Bascur, Rodrigo Costas, Suzan Verberne. 2024. Use of diverse data sources to control which topics emerge in a science map. arXiv. https://doi.org/10.48550/arXiv.2412.07550. [18]

users if their topics of interest do not align with the topic bias of the map, because then their topics would not be well represented by the map. For example, a science map user that wishes to find research about a given country will find no or few clusters about this country, leading to the wrong conclusion that there is little research about this country.

In the current paper, we explore whether the topic bias of a map can be adjusted by using different data sources to connect documents in the networks. In particular, we aim to identify combinations of sources and kinds of topics that show promise for achieving better clustering than traditional science map sources. This means that, rather than trying to outperform traditional science map networks across all topics, we focus on discovering cases where alternative sources provide complementary information that improves clustering for specific kinds of topics. This approach acknowledges that it is unrealistic to expect every topic to achieve high clustering effectiveness simultaneously, and instead seeks to offer science map users more targeted options depending on their topic of interest. In the example mentioned in the prior paragraph, a science map user interested in research about a given country could benefit from selecting a data source better suited to generating clusters about geographical locations.

The reason why we attempt to find effective combinations of sources and kinds of topics is that different sources contain different information about scientific content. For example, science maps that use patents as sources are likely to be more focused on technology than science maps that use text similarity. In this example, even if the science map based on patents has lower clustering effectiveness for all topics, its focus on technology could potentially be used in combination with a science map from a traditional source to increase the clustering effectiveness of technological topics, even if it diminishes the clustering quality for other kinds of topics.

The traditional data sources used to create science maps are citation links and text similarity, where connections are derived directly from the documents themselves. In this paper, we use the term data source to refer to any structured source of information used to connect academic documents. To achieve our goal, we explore other, non-traditional data sources. Most of our non-traditional data sources create networks where two or more academic documents are connected with an element external to the document (e.g. a patent that cites two documents), and for this reason we refer to these sources as external sources. Our topics are based on MeSH terms, and we group the topics into topic categories to facilitate our analysis. We measure the topic bias of a network as how well a topic is clustered (i.e. clustering effectiveness) over several clustering solutions, each of them with different cluster sizes. Each of these clustering solutions is analogous to a very simple science map. We use the topic bias of text similarity networks as our reference to compare how the topic bias changes in other networks.

Our research question is: Which topic categories benefit from using each external source? We operationalize this benefit in two ways: First, if the clustering effectiveness of the topic category in the network of the external source is higher than the effectiveness of the same topic category in the text similarity network. Second, if a topic category ranks among the higher-performing categories in clustering effectiveness within the external source, but not within the text similarity network. We will consider both operationalizations to address our research question, but give more importance to the first one because it serves the needs of science map users more directly.

Our contributions are: (1) We present an expanded and improved analysis method for evaluating the clustering effectiveness of a topic; (2) With this method, we provide a large-scale analysis of eight different sources (two traditional and six external), twenty one networks of up to four million documents, nearly three thousand clustering solutions, and seventeen topic categories, each one usually composed of between fifty and three hundred topics (values vary between networks); (3) With this analysis, we show that topic bias can be changed using external sources, and also which topics categories are favored for each of the external source. This knowledge expands the customization options of science maps.

## 5.2 Background

In this section we explore several topics related to our paper, provide literature examples for each of them, and explore how our paper relates to the most relevant ones.

### 5.2.1 Interaction of academic documents with non-academic elements

Traditionally, policy makers analyze scientific production to evaluate scientific impact, but they also are interested in evaluating its societal, technological and policy-making impact. For societal impact, the impact of publications on social media has been suggested as a proxy [172], and we highlight the company Altmetric [5, 53, 59], which collects mentions to academic documents online, including social media. For technological impact, patents are used [113]. Policy-making impact is a more recent field of study, and we highlight the company Overton [60, 150], which collects ample datasets of policy documents and their references [53]. We also highlight the company Dimensions [84], which collects the connections of academic documents to citations, clinical trials, patents, policy documents, grants and datasets.

### 5.2.2 Science maps based on diverse sources

Science maps of academic documents typically use networks of citation links or text similarity [165], but both Janssens, Glänzel, and De Moor [91] and Ahlgren et al. [4] proposed networks that combine both citation links and text similarity. Also, Costas, de Rijcke and Marres [45] proposed a conceptual framework for analyzing the interaction between documents and social media by creating networks of co-occurrence. Their framework is our source of inspiration for using external sources to improve science maps and also for how we build the networks of external sources. The main difference between their networks and our networks is that in their networks co-occurrence is explicitly included in the weight of the edges, while in our networks it is implicit by building the network with both the documents and the elements where the documents co-occur, an approach similarly to the work of Yun, Ahn and Lee [182].

An alternative method to create science maps is to create a network where the clusters are not made of academic documents, so to obtain a different perspective on the academic data. Keywords can be used to identify the topics within a collection of documents, connecting the keywords by the documents where they co-occur [103]. This has a slightly different functionality from identifying topics using document clusters, like to study the evolution of topics over time [167]. Authors can be used to identify scientific collaborations, connecting the authors either by their co-authorships [120] or their citations [166]. Patents can be used to identify technological developments, connecting the patents by their cited documents [102]. By their nature, networks of elements that co-occur with academic documents can be turned into networks of documents that co-occur with these elements. For example, Tang and Colavizza [179] created two networks using the same data, one of documents cited by the same Wikipedia article, and one of Wikipedia articles citing the same document. In this example, the co-occurrences where explicit, but Carusi and Bianchi [34] created a bipartite network of authors and journals where the co-occurrences were implicit. This allowed them to create clusters for both the authors and the journals using the same network with a method they called co-clustering. In our paper the external source networks are also bipartite, but our methodology will only focus on clustering the academic documents, not the external source elements.

### 5.2.3 Criticisms to maps of science

There are several criticisms of the capacity of science maps to represent topics. Gläser [64] reported that expert based evaluation of maps is usually inconclusive. Held, Laudel and Gläser [78] found that the science maps were unable to have both at the same time one topic per cluster and one cluster per topic. Held and Velden [76] found that clusters represent individual species instead of a biological field. Hric, Darst and Fortunato [86] made a strong criticism of the capacity of any kind of clustering algorithm in any kind of network to create clusters where all the cluster nodes belong to a given category. Because of the failure of science maps to properly cluster all topics, topic wise evaluation of science maps aims to make a more granular evaluation of the clustering and identify which topics get more effectively clustered, instead of making an overall statement about the quality of the map. This kind of evaluation has been sparsely explored by the literature. As far as we know, beyond our prior work [19], the only topical analyses that exist are the expert based evaluations of science maps and, to a lesser extent, the exploration of the epistemic function of intra- and inter-cluster citations performed by Seitz et al. [141].

### 5.2.4 Comparing clustering solutions of different networks

Different networks generate different science maps, and there have been several attempts to compare the clustering solutions of different networks. Xu et al. [178] identified overlapping communities between the clusters of two networks with the same nodes. Xie and Waltman [177] did something similar, but using topic modeling instead of text similarity networks. Šubelj, Van Eck and Waltman [148] evaluated the quality of the clusters generated by different clustering algorithms from the same network. Their method evaluated if the topics of the clusters correspond to the topics of the field experts, and also evaluated attributes of the clustering, like clustering stability, computing time, and cluster size. Waltman et al. [165] compared clustering solutions from different networks with the same nodes using an additional network as reference to calculate the accuracy of the clusters. For an example that does not use clustering, Ba and Liang [11] identified overlapping edges between two networks with the same nodes. In our prior work [19], we compared the clustering effectiveness per topic by evaluating the extent to which topic documents are in few clusters and the extent to which these same clusters only contain topic documents. In the current paper we refine this method so its results are easier to interpret.

## 5.3 Methods

In this section we describe how we obtained and cleaned the data, created the networks and clusters, evaluated the clustering effectiveness, and compared the topic categories.

### 5.3.1 Core academic documents

This is the set of documents that we used in the evaluation of clustering effectiveness, and each network has a different subset of these documents depending on the data available for each external source. We selected all Web of Science documents from the CWTS local database published between the years 2016 and 2019 that have a PubMed id (which is necessary to have MeSH terms) and that have a noun phrase in the title or abstract sections. The latter condition was added to have high quality text similarity networks, and the noun phrases were identified using the method developed by Waltman and van Eck [164]. We chose this range of years so as to have enough connections between the documents and the external source elements, especially with patents because they take multiple years to accumulate, and also because in these years Twitter became popular for sharing academic documents while not being the years of the Coronavirus pandemic. The external source elements are not limited to this period and instead go up to the year 2023. For example, a patent published in 2023 may cite a document from 2019. The time gap between social media posts and the documents they link to tends to be shorter than for other sources. In total, our core set contains 4,142,511 documents.

### 5.3.2 External sources networks

The external source networks are built the following way: For each external source, we first define what the nodes of this source mean (e.g. academic document authors, facebook users, etc...), which we will refer to as the external source "elements". Then we select core academic documents and external elements that we will use in the network, such that all the documents are connected to at least one external element and all the external elements are connected to at least two documents. We use the "at least two documents" threshold so that we do not have documents without any indirect connections with other documents (there are no direct connections between documents). Then we create a network with these documents and external elements where the edges that connect them are undirected and have weight value 1, the document nodes have weight value 1 and the external element nodes have weight value 0. We give this weight value to the external element nodes so that the clustering algorithm does not take these nodes into account when calculating the quality of a cluster. We will refer to these networks as the "Pure" networks of an external source, to distinguish them from the mixed and the text similarity networks of an external source (described in Section 5.3.3). It is worth mentioning that this network creation design creates a bipartite network (only document to external element edges), while in science mapping literature it is more common to represent these relations as a co-occurrence network (only document to document edges with no external element nodes, and the weight value of the edge is the number of external elements in common between the documents). We use bipartite networks because they represent these relations with more computational efficiency than co-occurrence networks. This happens because, even as the bipartite network has more nodes because it must also represent the external elements, the number of edges is much lower because the co-occurrences are not represented explicitly with document-to-document edges.

We used the following external sources. All databases are the local version from CWTS, version year 2023:

**Documents authors (AUTHOR):** The external source elements are the authors of academic documents, and the connections are to these documents. The data comes from the disambiguated authors database of CWTS [54]. This network has 3,977,303 core academic documents, 2,710,012 external source elements and 19,820,564 edges.

**Facebook users (FACEBOOK):** The external source elements are the Facebook users (i.e. accounts), and the connections are to the documents they have posted web links to. The data comes from the Altmetric [5] Facebook database. This network has 596,783 core academic documents, 44,811 external source elements and 1,231,887 edges.

**Twitter users (TWUSER):** The external source elements are the Twitter users (i.e. accounts), and the connections are to the documents that their tweets have web links to. The data comes from the Altmetric [5] Twitter database. This network has 2,364,304 core academic documents, 1,495,275 external source elements and 27,981,494 edges.

**Twitter conversations (TWCONV):** The external source elements are the Twitter conversations, and the connections are to the documents that its tweets have web links to. A Twitter conversation is an original (non-reply) tweet plus all the tweets that directly or indirectly reply to it. The data comes from the Altmetric [5] Twitter database. This network has 227,212 core academic documents, 493,049 external source elements and 1,175,624 edges. Notice that this network is substantially smaller than the TWUSER network, even though both are created from the same database. This is because many documents are connected by the same Twitter user, but fewer are connected by the same Twitter conversation.

**Patents families (PATENT):** The external source elements are patent families, and the connections are to the documents cited by the patents of the patent family. A patent family is made up of an initially submitted patent, plus derivative patents (like updates or new applications) and versions of the patent submitted in different countries. The data comes from the PATSTAT database [93] and we only use invention patents. This network has 98,278 core academic documents, 41,714 external source elements and 175,693 edges.

**Policy documents (POLICY):** The external source elements are policy documents, and the connections are to the documents cited by the policy documents. A policy document is a document written primarily for policy makers, and includes documents such as memos and guidelines from governments and think tanks. The data comes from the Overton database [150]. This network has

311,867 core academic documents, 64,951 external source elements and 651,099 edges.

### 5.3.3 Text similarity networks

We use the topic bias of text similarity networks in our experiments as a reference to compare how the topic bias changes in other networks. We chose this source because it is traditionally used for the creation of science maps and also because it is less computationally demanding to create and cluster than the citation network, which is relevant because we created a reference network for each external source. The method to measure text similarity was the cosine similarity between the embedding of the text of two documents. The text of a document is its concatenated title and abstract, and the embedding is extracted using the Python implementation of Sentence BERT [132] with the "allenai-specter" model [43], which is a model specifically trained with scientific literature. These methods have already been used for scientometric tasks. For example, OpenAlex trained their academic topic classifier using Sentence BERT and the clusters of a science map [123], while Woo and Walsh [174] used the same model as us to measure the text similarity between academic documents.

For each external source, we create a text similarity network that contains the same academic core documents as the Pure network, which we will refer to as the "BERT" network, and we also create a network that combines both networks, which we will refer to as "Mixed" network. To create the BERT network of a source we first make the academic documents into nodes with weight value 1. Then, we calculate the text similarity between all pairs of documents and only keep the 20 highest pairs per document. These values become the weights of the undirected edges between the nodes, and if there are two edges between two nodes then we merge them and sum their weights. Finally, we multiply all the edge weight values by a factor such that the sum of all edge weight values in a network is the same for the BERT and the Pure networks. To create the Mixed network of a source we use the Pure network and add to it the edges from the BERT network. The purpose of the step where we multiply the edge weight values by a factor is to bring this network to the same magnitude as the Pure network, which has two goals: To make the edges that came from the BERT and Pure network have the same magnitude of influence in the edges of the Mixed network, and to use the same clustering Resolution values for the BERT and Pure networks, which is just convenient.

### 5.3.4 Citation network

There are not many science maps studies published using Sentence BERT for text similarity because it is a recently developed method, making our results difficult to compare to the literature. To solve this, we also evaluated the topic bias of a network that is built based on a method well researched in the literature and presented it next to the other external source networks. This well published method is the extended direct citation [165], which is a citation network that includes connections to academic documents that are not part of the core academic documents. The Pure citation network includes all the core academic documents as nodes with weight value 1 and the citations between each other as undirected edges with weight value 1. It also includes the non-core documents from Web of Science that have citation links to at least two core academic documents as nodes with weight value 0, and these links as undirected edges with weight value 1. These non-core documents are documents from outside the time period or that do not have a PubMed id, which means they are likely not about biomedical topics. This network has 4,142,511 core academic documents, 18,960,516 non-core academic documents and 217,907,980 edges. The Mixed and BERT citation networks were created the same way as for the external sources (the BERT network uses only the core academic documents).

We considered creating a citation network for the documents in each external source, just like we did for the text similarity network, because both are typically used in science mapping. However, we ultimately decided to only do this with the text similarity network for two reasons. First, due to the external source documents being a subset of the full core set, some of them would lose many of their citation links when restricted to this smaller subset. This would reduce the quality of the resulting clusters. This issue does not affect text similarity because it can be calculated between any pair of documents. Second, citation networks are significantly larger than text similarity networks due to the high number of additional nodes that come from the extended citation, making the clustering process much slower. Additionally, even when using a smaller set of core documents in the external source networks, the size of the citation network does not decrease proportionally. This happens because many of the removed core documents still appear in the network as non-core document nodes, as they tend to cite at least two documents from the smaller set due to the close publication years.

### 5.3.5 Clustering

To cluster we used the Leiden algorithm [153], which is typically used in science maps. This algorithm requires the user to set a parameter, the “Resolution”, which has an effect on the size of the clusters (higher Resolution, smaller clusters). We clustered each network several times using a wide range of Resolution values, using a different value each time. We decided on the Resolution values range on a network wise basis, and our criteria for this range was for the highest value to create a clustering solution where most clusters have only one node, and for the lowest value to create a clustering solution where most of the nodes belong to a single cluster. We clustered a number of Resolution values that allowed us to keep the running time manageable (between 70 and 140 Resolution values per network), using the Python implementation of the library Igraph [47] and the Leiden algorithm. All the clustering solutions are used during the evaluations and comparisons.

### 5.3.6 Topics and topic categories

Our topics are the tree nodes in the MeSH hierarchical tree of MeSH terms, and the topic documents of a given topic are the documents labeled with the tree node of a topic. MeSH terms are a controlled vocabulary thesaurus from the National Library of Medicine (NLM) used for indexing PubMed, and are semi-automatically annotated to documents by the NLM [117]. We use MeSH terms instead of other alternatives because of their extensive system of hierarchical topics, high number of annotated documents, and high quality of annotations. The MeSH terms are organized in a hierarchical tree where almost each MeSH term maps to one or more nodes in the tree, but each tree node maps to a single MeSH term. The tree is composed of 16 branches, and the tree nodes in the lower levels are subtopics of the tree nodes in the higher levels. We refer to a tree node using its MeSH term name followed by their tree node identity (e.g. *Head [A01.456]*). The reason why we base our topics on the tree nodes of the MeSH terms instead of just using the MeSH terms themselves is to facilitate the expansion and filtering of topics in the next steps of the methodology (see below). We obtained the MeSH terms annotated for each document, plus the metadata of the MeSH terms themselves, including their tree nodes, from the in-house CWTS database of PubMed and MeSH (version from 2024).

Our topic categories are the MeSH tree branches, and all the tree nodes in the branch are topics that belong to the topic category. We use branches as topic categories because they are epistemic categories (e.g. organisms), which are the kind categories commonly used for topical analysis of clusters [19, 141]. There are 3 branches that we decided to, instead of using them as topic categories, use their highest level tree nodes as topic categories, because we think these tree nodes work better than their branches as topic categories. The branches that we replaced with their higher level tree nodes are *Disciplines and Occupations [H]*, *Anthropology, Education, Sociology, and Social Phenomena [I]* and *Technology, Industry, and Agriculture [J]*. We also removed the following topic categories due to having too few topics: *Humanities [K]*, *Publication Characteristics [V]*, *Human Activities [I03]*, and *Non-Medical Public and Private Facilities [J03]*. In the end, we used the 17 topic categories in Table 5.1.

To have good topics, we would like each topic to be annotated on all the documents related to it, but the NLM typically only annotates up to fifteen MeSH terms per document, which means that the more generic MeSH terms are not annotated. To fix this, we expanded the topics annotated on a document using the already annotated MeSH terms and the MeSH tree. We transformed each

Table 5.1: List of topic categories used in the current paper.

| **Topic Categories** |
|---|
| Anatomy [A] |
| Organisms [B] |
| Diseases [C] |
| Chemicals and Drugs [D] |
| Analytical, Diagnostic and Therapeutic Techniques, and Equipment [E] |
| Psychiatry and Psychology [F] |
| Phenomena and Processes [G] |
| Natural Science Disciplines [H01] |
| Health Occupations [H02] |
| Social Sciences [I01] |
| Education [I02] |
| Technology, Industry, and Agriculture [J01] |
| Food and Beverages [J02] |
| Information Science [L] |
| Named Groups [M] |
| Health Care [N] |
| Geographicals [Z] |

of the MeSH terms into all of their corresponding MeSH tree nodes, and then we added all the MeSH tree nodes upstream in the MeSH tree from the current MeSH tree nodes. For example, if a document had the MeSH term *Scalp*, we transformed this MeSH term into its tree node version (*Scalp [A01.456.810]*), and added the upstream tree nodes (*Head [A01.456]*, *Body Regions [A01]*) to the document. This MeSH term expansion is based on the "MeSH term explosion" feature of the PubMed online search interface.

To improve the reliability of our evaluation we filter our topics. We do this filtering process for each external source because they use different sets of core academic documents. Our first filter criterion is by topic size (i.e. number of documents with the topic) because the size of a topic can affect its clustering effectiveness. We group the topics by size into Size bins, which go from a value (excluding it) to double that value (including it), starting at 40 (e.g. 41-80, 81-160, 161-320, ... $[X+1]$-$[2X]$). We use 40 for reasons explained in Section 5.3.7.1. We filter out the Size bins that have less than half the number of topics than the Size bin with most topics, and also filter out the topics that belonged to these filtered out Size bins. The Size bins that we keep per source are shown in Table 5.2.

Table 5.2: Size bins per source after filtering.

| **Source** | **Size Bins** |
|---|---|
| Patents families | 41-80; 81-160; 161-320 |
| Policy documents | 41-80; 81-160; 161-320 |
| Facebook users | 41-80; 81-160; 161-320; 321-640 |
| Twitter conversations | 41-80; 81-160; 161-320; 321-640 |
| Twitter users | 81-160; 161-320; 321-640; 641-1,280 |
| Documents authors | 161-320; 321-640; 641-1,280; 1,281-2,560 |
| Citations | 161-320; 321-640; 641-1,280; 1,281-2,560 |

Our second filter criterion is redundancy (i.e. two topics share a substantial number of documents) because it can distort our results. To filter by redundancy, we first identify the topics within the same topic category that are redundant with each other. We define two topics as being redundant if they have a Jaccard similarity of 0.5 or higher (calculated from their number of shared documents).

We group the redundant topics using the agglomerative hierarchical clustering algorithm with the Complete Linkage method [137] and Jaccard distance, with 0.5 as threshold. Then, we filter out each but the smallest topic from each group, which in our experience tends to also be the topic that best describes the group. For example, if there is a group of redundant topics made up of *Canidae [B01.050.150.900.649.313.750.250.216]* and *Dogs [B01.050.150.900.649.313.750.250.216.200]*, we believe that this group is better described by the latter than the former. In cases where a group had more than one smallest topic, we selected the one with the tree node at the lowest level in the tree. If there is more than one at this level, we select one using a deterministic random process. After filtering topics, we also filter the topic categories that contain too few topics in any Size bin. We choose this threshold manually per external source, but it is always at least between 5 and 10 topics. It is worth mentioning that in our prior work [19] we defined two topics as being redundant if they had Jaccard similarity 0.9 or higher, so in the current paper we are being substantially stricter at ensuring the quality of the data.

### 5.3.7 Evaluation

#### 5.3.7.1 Clustering effectiveness

To find out which topics are better represented by the clustering of the networks, we use the concept of clustering effectiveness that we introduced in our prior work [19]. The unit to measure the clustering effectiveness is "Purity", which is, for a set of selected clusters, which fraction of their documents belong to a given topic. In mathematical terms, Purity is defined as:

$$Purity = \frac{\sum_{i=1}^{N} |D_i \cap D_M|}{\sum_{i=1}^{N} |D_i|} \tag{5.1}$$

Here, $N$ denotes the number of selected clusters, $D_i$ denotes the documents in selected cluster $i$ and $D_M$ denotes the topic documents of the topic. The higher Purity, the more effective the clustering. Purity is bounded between values 0 and 1, with Purity value 1 meaning that the selected clusters only contain topic documents. We calculate Purity for each clustering solution and topic, but instead of selecting all the clusters that contain topic documents to calculate Purity, we only select a subset of these clusters. To do this, we sort all the clusters that contain topic documents from the highest to the lowest number of topic documents, with ties won by the smallest cluster. Then, we choose the threshold of the minimum number of topic documents that we want the set of selected clusters to contain, and then select clusters in the sorted order until we reach this threshold. We call this value Coverage, and it is a fraction of the total number of topic documents. In our paper we calculate Purity for three Coverage values: 0.25, 0.50 and 0.75. We only compare Purity values calculated using the same Coverage value. In reference to Section 5.3.6, the reason why Size bins start at 40 is because at Coverage 0.25 the value of the threshold is only 10 documents, which we set as the minimum number to have a meaningful academic topic.

In our concept of clustering effectiveness, the number of selected clusters (NSC) also plays a role. In a science map, finding clusters related to a topic requires effort, so the smaller the NSC, the higher the cluster effectiveness. Also, a high NSC is correlated with smaller clusters, which itself is correlated with higher Purity because smaller clusters allow a more fine selection of the clusters. For example, if all clusters in a clustering solution are size 1, then the value of Purity is also 1 because all the selected clusters contain only topic documents. To control for the effect of NSC over Purity, we only compare Purity values when they have the same NSC.

#### 5.3.7.2 Topic Purity profiles

In our research question, we operationalized the concept of "benefit" in two ways. The first operationalisation was if the clustering effectiveness of the topic category in the external source (either the Pure or Mixed network) is higher than the same topic category in text similarity (the BERT

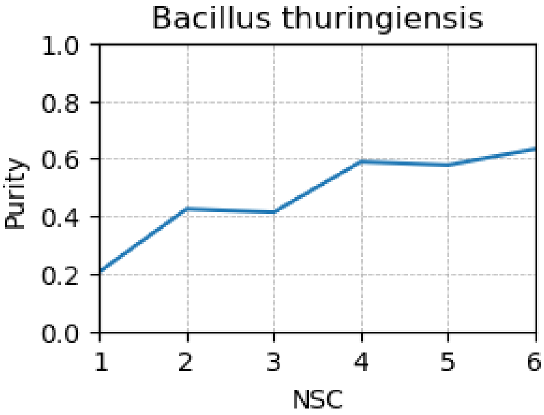


Figure 5.1: Example of a Purity profile. This is a line plot of the Purity profile of the topic *Bacillus thuringiensis [B03.510.460.410.158.218.800]* for the Policy documents BERT network calculated using Coverage 0.50. This topic has 60 topic documents among the core documents used by the Policy networks, which for this Coverage value means that the Purity is calculated after selecting clusters that contain at least 30 topic documents. So for example, if we assume that the selected clusters contain exactly 30 topic documents, from the figure we can say that at different Resolution values the network can place 30 out the 60 topic documents in one cluster containing 150 documents (30/0.2), two clusters containing 75 documents (30/0.4), and four clusters containing 50 documents (30/0.6). Using lower Coverage values or topics with more topic documents tends to achieve higher Purity at the highest NSC value.

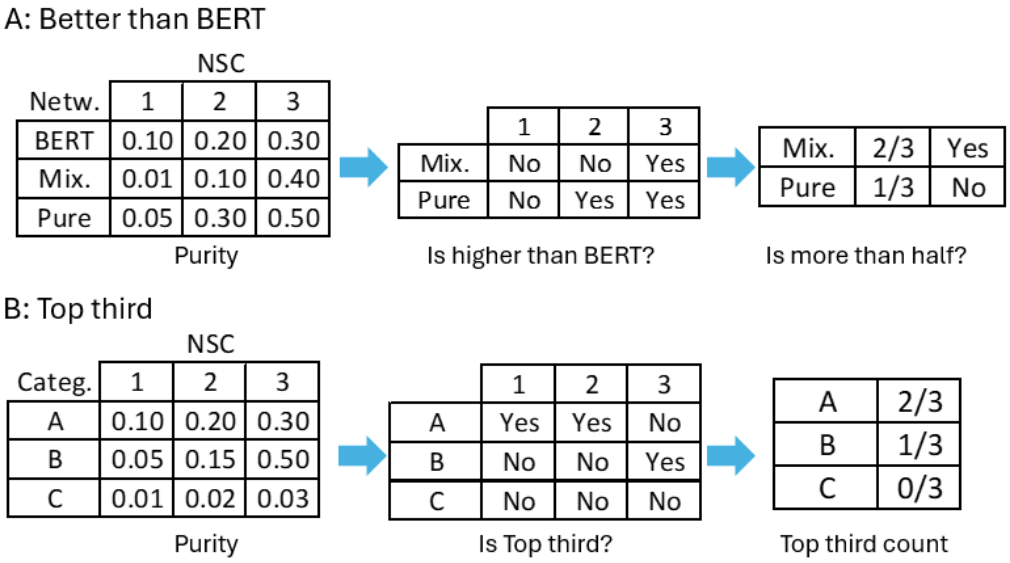
network). We answer this question by comparing the clustering effectiveness of each topic between these networks. We represent the clustering effectiveness of a topic for a given network as a series of NSC–Purity value pairs that we will refer to as the ”topic Purity profile”. The NSC values are a consecutive sequence of integers that go from 1 to $N$, and $N$ is:

$$N = \lfloor \frac{S * Cov}{5} \rfloor \tag{5.2}$$

Here, $S$ is the size of the topic, $Cov$ is the coverage value, and the function $\lfloor x \rfloor$ means rounded down to the nearest integer. Therefore, the number of NSC values in a Purity profile depends on the size of the topic. The denominator 5 ensures that, at the highest NSC value, the average number of topic documents per selected cluster is at least 5, so to limit the NSC to a value that is meaningful in a science map context. The first value of NSC is 1 because it is the minimum number of selected clusters.

For each NSC value, we assign the highest available Purity value among clustering solutions with the same NSC. If there is no clustering solution with NSC value 1, we assign to it Purity value 0. If there is no clustering solution with any of the other NSC values, we estimate its Purity value by linear interpolation between the Purity values of the two nearest NSC values with known Purity. If necessary, we interpolate using the Purity value of NSC values higher than $N$. An example of a topic's Purity profile is shown in Figure 5.1.

We say that a topic has higher clustering effectiveness in one network than in another if more than half of its NSC values have higher Purity in one network than in the other. Figure 5.2A shows an example diagram of how we calculate this. For each topic category, we calculate the fraction of their topics that have higher clustering effectiveness in the Mixed or Pure network than in the BERT network. We refer to this value as “absolute Purity difference” of this topic category, and it answers the first operationalisation of our research question. For example, if the absolute Purity difference of a topic category in the Pure network of an external source is 0.25, it means that a quarter of its



A: Better than BERT

NSC

| Netw. | 1 | 2 | 3 |
|---|---|---|---|
| BERT | 0.10 | 0.20 | 0.30 |
| Mix. | 0.01 | 0.10 | 0.40 |
| Pure | 0.05 | 0.30 | 0.50 |

Purity

| | 1 | 2 | 3 |
|---|---|---|---|
| Mix. | No | No | Yes |
| Pure | No | Yes | Yes |

Is higher than BERT?

| | | |
|---|---|---|
| Mix. | 2/3 | Yes |
| Pure | 1/3 | No |

Is more than half?

B: Top third

NSC

| Categ. | 1 | 2 | 3 |
|---|---|---|---|
| A | 0.10 | 0.20 | 0.30 |
| B | 0.05 | 0.15 | 0.50 |
| C | 0.01 | 0.02 | 0.03 |

Purity

| | 1 | 2 | 3 |
|---|---|---|---|
| A | Yes | Yes | No |
| B | No | No | Yes |
| C | No | No | No |

Is Top third?

| | |
|---|---|
| A | 2/3 |
| B | 1/3 |
| C | 0/3 |

Top third count

Figure 5.2: Diagram on the representation of results. A: How to calculate from topic Purity profiles if a topic has higher clustering effectiveness than BERT in the Pure or the Mixed network. In this example, a topic has higher Purity than BERT for the Mixed network, but not so for the Pure network. B: How to calculate from topic category Purity profiles the number of NSC that a topic category is in the top third Purity of a network. In this example, the topic categories A, B and C achieve a top third count of 0.7, 0.3 and 0, respectively.

topics have higher Purity in the Pure network than the BERT network.

#### 5.3.7.3 Topic category Purity profiles

The second operationalization of ”benefit” is if a topic category ranks among the higher-performing categories in clustering effectiveness within the external source (either the Pure or Mixed network), but not within the text similarity network (the BERT network). We answer this question by comparing the clustering effectiveness of all topic categories within each network.

The topic Purity profile defined in Section 5.3.7.2 represents the clustering effectiveness of individual topics in a given network. However, in the current section we need to define a representation at the level of topic categories. To achieve this, we introduce the concept of ”topic category Purity profile”. We create a different Purity profile for each Size bin, because higher Size bins require higher NSC values and achieve higher Purity. Without separating by Size bin, comparisons between topic categories would be affected by which topic category has larger topics.

The Purity profile is a series of NSC–Purity value pairs, where the NSC values are a consecutive sequence of integers that go from 1 to $N$. $N$ is calculated the same as in Equation 5.2, but $S$ is not the size of the topic but the size of the Size bin, which we define as the average between the lower and upper bound of the Size bin (e.g. for the Size bin 41-80, $S = 60$, and if $Cov = 0.25$, then $N = 3$).

To assign Purity values to the NSC values, we do the following: For each clustering solution, we average the Purity values and the NSC values of all the topics that belong to the topic category and Size bin. Then, for each NSC in the Purity profile, we assign a Purity value using the same interpolation method described in Section 5.3.7.2, using the averaged NSC–Purity pairs obtained from the clustering solutions. It is worth mentioning that we also considered using topic category Purity profiles instead of topic Purity profiles for the first operationalization of benefit, but we found that the results from this approach provided us with less nuanced information than the one we ultimately used.

To answer our operationalization of benefit, we first identify which topic categories are among the higher-performing categories in each network. We take all the topic category Purity profiles for a given network within the same Size bin, and for each NSC value, we identify the topic categories that rank among the top third based on Purity. We then calculate, for each topic category, the fraction of NSC values for which it is among the top third. Figure 5.2B shows an example diagram of how we calculate this value. This fraction, averaged across all Size bins of the topic category, is referred to as the "top third count".

The top third count represents the tendency of a topic category to be among the higher-performing topic categories of a network. For example, if the top third count of a topic category in a network is 0.25, it means that, on average across the Size bins, it is among the top third highest Purity topic categories for a quarter of the NSC. We define the top group of topic categories in relative terms (as a third) instead of absolute terms (e.g. top three) because different external sources have a different number of topic categories due to the topic category filtering in Section 5.3.6.

Finally, we compare the top third count of each topic category between the Pure or Mixed network and the BERT network by subtraction (e.g. Pure top third count minus BERT top third count). We refer to this value as the "relative Purity difference", which is used to answer the second operationalization of our research question.

### 5.3.8 Summary of methods

The methodology consists of two parts: The measurement of clustering effectiveness, and the evaluation of clustering effectiveness. We group the relevant variables in brackets at each step to improve clarity and readability.

The steps of the measurement are:

1. For each [external source], we select a subset of the core documents.

   1.a. We map these documents to topics. The topics that are too small and the topic categories with too few topics are discarded from the experiment.

   1.b. We create a Pure, Mixed, and BERT network with these documents.

2. For each [external source and network], we generate multiple clustering solutions using different Resolution values.

3. For each [external source, network, clustering solution and topic], we select the relevant clusters using each of the different Coverage values.

4. For each [external source, network, clustering solution, topic and Coverage value], we compute two metrics for the selected clusters: NSC and Purity.

The evaluation consists of two tracks: One for absolute Purity difference, and one for relative Purity difference.

The steps for calculating the absolute Purity difference are:

1. For each [external source, network, topic and Coverage value], we create a topic Purity profile using the NSC and Purity values from all the clustering solutions. The topic Purity profiles from the same [external source, topic and Coverage value] share the same NSC values, which enables comparison.

2. For each [external source, topic and Coverage value], and for the Pure and Mixed networks, we compute the fraction of NSC values where the Purity is higher than in the BERT network. If this occurs for more than half of the NSC values, we label the topic as having better clustering effectiveness in that network than in the BERT network (Figure 5.2A).

3. For each [external source, topic category and Coverage value], and for the Pure and Mixed networks, we compute the fraction of topics in the topic category that had higher clustering effectiveness. This final value is the absolute Purity difference.

The steps for calculating the relative Purity difference are:

1. For each [external source, network, Size bin, clustering solution, topic category and Coverage value], we calculate the average NSC and Purity values across all the topics of the topic category within the same Size bin.

2. For each [external source, network, Size bin, topic category and Coverage value], we create a topic category Purity profile using the averaged NSC and Purity values from all clustering solutions. The topic category Purity profiles from the same [Size bin and Coverage value] share the same NSC values, which enables comparison.

3. For [external source, network, NSC, Size bin and Coverage value], we sort topic categories by Purity (highest first) at that NSC, and record which topic categories are in the top third of the ranking.

4. For each [external source, network, Size bin, topic category and Coverage value], we compute the fraction of NSC values where the topic category appears in the top third (Figure 5.2B).

5. For each [external source, network, topic category and Coverage value], we average these values across all Size bins. This average is the top third count of the topic category.

6. For each [external source, topic category and Coverage value], and for the Pure and Mixed networks, we report the difference between that network and the BERT network in the top third count. This final value is the relative Purity difference.

## 5.4 Results

In this section, we present our results, discuss the performance of each external source, and explore in depth the cases with the best performance. From this point on, we refer to specific networks of an external source using the following prefixes: “b” for BERT, “m” for Mixed, and “p” for Pure. For example, “mTwconv” refers to the Mixed network of the Twitter conversations. We avoid exploring the following results in depth:

1. Topic category *Organisms [B]*: Most external sources, including citations, outperform BERT on this category, suggesting that BERT performs particularly poorly here.

2. Citation networks: While included for comparison, our focus is on external sources. The citation network serves mainly to connect our findings to prior work on citation-based science maps.

3. Coverage values: The three tested values produced similar results, with only a few exceptions.

The results of our experiments are presented in detail in Table 5.3 and summarized in Table 5.4. The summary transforms the top third counts into relative Purity differences, reports only the highest absolute and relative Purity differences among the three Coverage values, and uses signs and colors instead of numerical values. In Table 5.5, we indicate which networks perform best per topic category, and by how much. We analyze these topic categories Purity profiles (examples shown in Figure 5.3) to assess whether they are “competitive”, meaning that their Purity values are close to or exceed those of BERT, and therefore might generate science maps of comparable quality. Finally, we include individual topic examples from some of these topic categories (Figure 5.4) to provide a more concrete illustration of our results.

Table 5.3: Detail of the results of each network. For each topic category, we show the top third count and the absolute Purity difference at each Coverage value. Zero values are omitted. Dots mean that the topic category was not included in the experiment due to having too few topics per Size bin, as explained in the filtering process.

| CITATION | Top third count | | | | | | | | | Absolute Purity Difference | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Network | BERT | | | Mixed | | | Pure | | | Mixed | | | Pure | | |
| Coverage | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 |
| Anatomy. | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.6 | 0.6 | 0.6 | 0.2 | 0.2 | 0.1 |
| Organisms. | 0.3 | 0.1 | 0.3 | 0.7 | 0.3 | 0.4 | 1.0 | 1.0 | 1.0 | 0.9 | 0.9 | 0.8 | 0.7 | 0.7 | 0.6 |
| Diseases. | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.7 | 0.8 | 0.8 | 0.2 | 0.3 | 0.3 |
| Chemicals . | | | | | | | 0.3 | 0.3 | | 0.8 | 0.8 | 0.6 | 0.6 | 0.6 | 0.5 |
| Analytical. | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.8 | 1.0 | 1.0 | 0.6 | 0.5 | 0.5 | 0.1 | 0.1 | 0.1 |
| Psychiatry. | 0.7 | 0.9 | 0.7 | 0.5 | 0.7 | 0.5 | | 0.1 | 0.4 | 0.7 | 0.7 | 0.6 | 0.2 | 0.2 | 0.2 |
| Phenomena . | | 0.2 | 0.4 | | 0.1 | 0.3 | | | 0.1 | 0.6 | 0.6 | 0.5 | 0.2 | 0.2 | 0.1 |
| Natural Sc. | | | | | | | | | | 0.5 | 0.3 | 0.3 | 0.2 | 0.1 | |
| Health Occ. | | | | | | | | | | 0.5 | 0.5 | 0.5 | 0.2 | 0.1 | 0.1 |
| Social Sci. | | | 0.1 | | | 0.2 | | | | 0.7 | 0.7 | 0.6 | 0.3 | 0.3 | 0.2 |
| Education. | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| Technology. | 0.6 | 0.5 | 0.2 | 0.4 | 0.5 | 0.2 | 0.3 | 0.2 | 0.3 | 0.6 | 0.7 | 0.6 | 0.2 | 0.3 | 0.1 |
| Food and B. | 0.4 | 0.3 | 0.3 | 0.4 | 0.3 | 0.3 | 0.5 | 0.4 | 0.3 | 0.6 | 0.7 | 0.7 | 0.2 | 0.2 | 0.2 |
| Informatio. | | | | | | | | | | 0.5 | 0.4 | 0.3 | 0.1 | 0.1 | |
| Named Grou. | | | | | | | 0.1 | | | 0.7 | 0.7 | 0.5 | 0.4 | 0.4 | 0.2 |
| Health Car. | | | | | | | | | | 0.6 | 0.6 | 0.4 | 0.2 | 0.2 | 0.1 |
| Geographic. | | | | | | | | | | 0.7 | 0.5 | 0.4 | 0.6 | 0.3 | 0.2 |

| TWCONV | Top third count | | | | | | | | | Absolute Purity Difference | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Network | BERT | | | Mixed | | | Pure | | | Mixed | | | Pure | | |
| Coverage | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 |
| Anatomy. | 1.0 | 1.0 | 0.9 | 0.9 | 0.9 | 0.9 | | | 0.1 | 0.3 | 0.2 | 0.2 | | | |
| Organisms. | 0.5 | 0.5 | 0.8 | 0.6 | 0.5 | 0.8 | 0.5 | 0.5 | 0.5 | 0.3 | 0.3 | 0.3 | | | |
| Diseases. | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.6 | 0.2 | 0.3 | 0.2 | 0.3 | 0.3 | | | |
| Chemicals . | | | | 0.1 | | | 0.2 | 0.1 | 0.1 | 0.4 | 0.3 | 0.3 | | | |
| Analytical. | 0.5 | 0.5 | 0.3 | 0.5 | 0.5 | 0.3 | 0.2 | | 0.1 | 0.2 | 0.2 | 0.2 | | | |
| Psychiatry. | 0.3 | 0.3 | 0.3 | 0.2 | 0.3 | 0.3 | 0.4 | 0.6 | 0.7 | 0.2 | 0.2 | 0.2 | | | |
| Phenomena . | | | | | 0.1 | | 0.2 | 0.2 | | 0.3 | 0.2 | 0.2 | | | |
| Natural Sc. | | | | | | | | 0.2 | 0.5 | 0.3 | 0.2 | 0.2 | | | |
| Health Occ. | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| Social Sci. | | | | | | | 0.7 | 1.0 | 0.9 | 0.4 | 0.4 | 0.3 | | | |
| Education. | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| Technology. | 0.7 | 0.7 | 0.6 | 0.7 | 0.7 | 0.6 | 0.4 | 0.4 | 0.2 | 0.3 | 0.3 | 0.3 | | | |
| Food and B. | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.9 | 0.4 | 0.3 | 0.5 | 0.1 | | 0.1 |
| Informatio. | | | | | | | | 0.1 | | 0.2 | 0.2 | 0.2 | | | |
| Named Grou. | | | | | | | 0.8 | 0.5 | 0.5 | 0.4 | 0.4 | 0.2 | 0.1 | 0.1 | |
| Health Car. | | | | | | | 0.1 | 0.1 | | 0.3 | 0.3 | 0.3 | | | |
| Geographic. | | | | | | | | 0.1 | 0.2 | 0.5 | 0.3 | 0.2 | 0.1 | | |

| AUTHOR | Top third count | | | | | | | | | Absolute Purity Difference | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Network | BERT | | | Mixed | | | Pure | | | Mixed | | | Pure | | |
| Coverage | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 |
| Anatomy. | 1.0 | 1.0 | 1.0 | 0.9 | 1.0 | 1.0 | | | | | | | | | |
| Organisms. | 0.3 | 0.1 | 0.3 | 0.8 | 0.7 | 0.9 | 1.0 | 1.0 | 1.0 | 0.1 | 0.1 | 0.1 | 0.2 | 0.1 | |
| Diseases. | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.9 | 0.8 | 0.5 | | | | | | |
| Chemicals . | | | | | | | | | | | | | | | |
| Analytical. | 1.0 | 1.0 | 1.0 | 0.9 | 1.0 | 1.0 | 0.3 | 0.2 | 0.1 | | | | | | |
| Psychiatry. | 0.7 | 0.8 | 0.4 | 0.6 | 0.4 | 0.4 | 0.1 | 0.3 | 0.7 | | | | | | |
| Phenomena . | | 0.1 | 0.6 | | 0.1 | 0.1 | | | | | | | | | |
| Natural Sc. | | | | | | | 0.1 | 0.1 | | | | | | | |
| Health Occ. | 0.1 | | | 0.2 | | | 0.9 | 1.0 | 0.9 | | | | | | |
| Social Sci. | | | 0.1 | | | 0.2 | | | | | | | | | |
| Education. | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| Technology. | 0.6 | 0.5 | 0.2 | 0.2 | 0.5 | 0.2 | | | | | | | | | |
| Food and B. | 0.4 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.1 | | | | | | | | |
| Informatio. | | | | | | | | | | | | | | | |
| Named Grou. | | | | | | | 0.6 | 0.6 | 0.7 | 0.1 | | | 0.1 | 0.1 | 0.1 |
| Health Car. | | | | | | | | | 0.1 | | | | | | |
| Geographic. | | | | | | | 1.0 | 1.0 | 1.0 | 0.4 | 0.1 | | 1.0 | 0.9 | 0.8 |

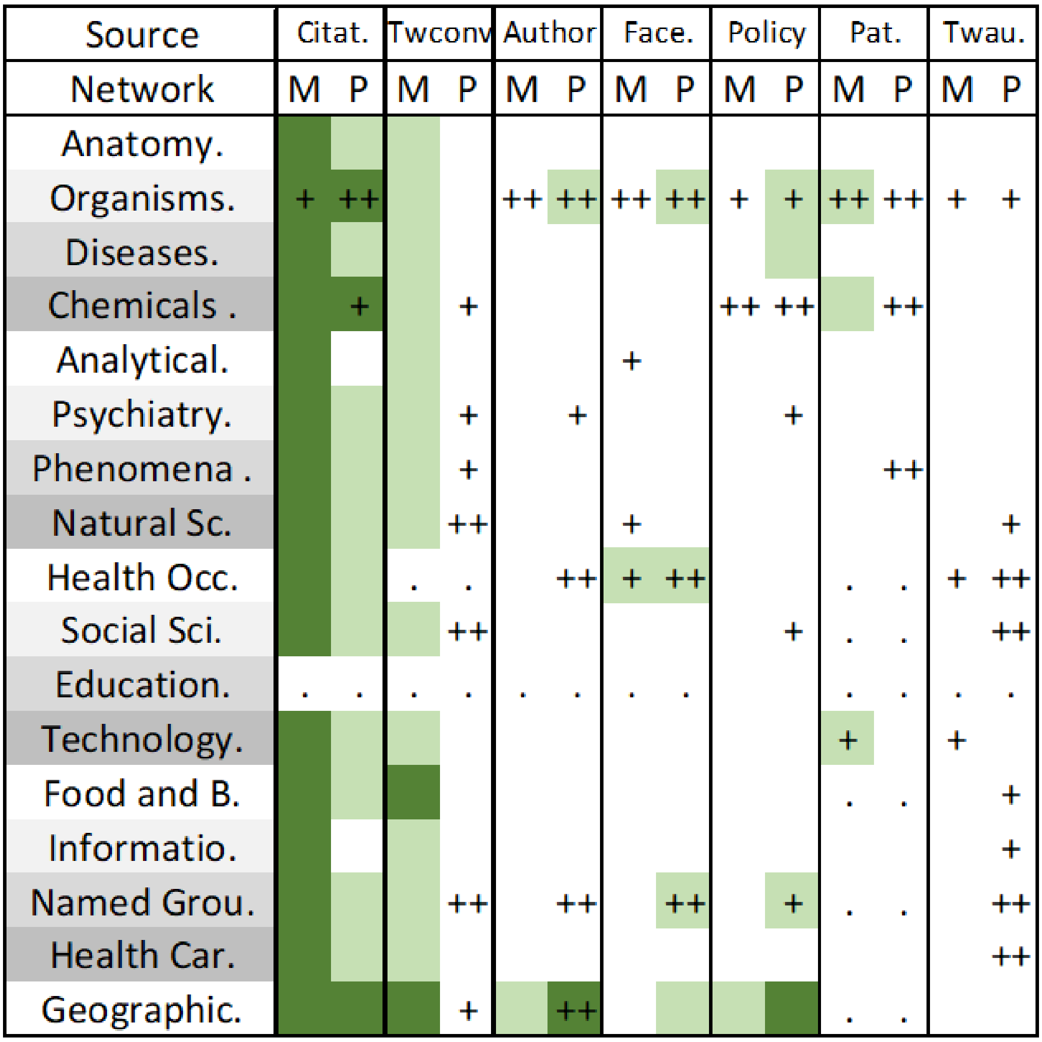
| FACEBOOK | Top third count | | | | | | | | | Absolute Purity Difference | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Network | BERT | | | Mixed | | | Pure | | | Mixed | | | Pure | | |
| Coverage | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 |
| Anatomy. | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.9 | 0.2 | 0.6 | 0.7 | | | | | | |
| Organisms. | 0.2 | 0.2 | 0.9 | 0.6 | 0.9 | 1.0 | 0.9 | 0.9 | 1.0 | 0.1 | 0.1 | 0.1 | 0.2 | 0.1 | 0.1 |
| Diseases. | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.9 | | | | | | |
| Chemicals . | | | | | | | | | | 0.1 | | | | | |
| Analytical. | 0.8 | 0.7 | 0.7 | 0.8 | 0.7 | 0.9 | 0.8 | 0.8 | 0.7 | | | | | | |
| Psychiatry. | 0.3 | 0.3 | 0.1 | 0.1 | | 0.1 | | | 0.1 | | | | | | |
| Phenomena . | 0.1 | 0.3 | 0.2 | | | | | | | | | | | | |
| Natural Sc. | 0.1 | 0.2 | 0.2 | 0.2 | 0.2 | 0.4 | 0.2 | 0.1 | 0.2 | 0.1 | | 0.1 | 0.1 | 0.1 | 0.1 |
| Health Occ. | 0.2 | | | 0.5 | 0.4 | 0.3 | 1.0 | 1.0 | 0.9 | 0.2 | 0.2 | 0.2 | 0.3 | 0.4 | 0.3 |
| Social Sci. | | | | | | | | | | | | | 0.1 | | |
| Education. | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| Technology. | 0.4 | 0.6 | 0.1 | 0.4 | 0.4 | 0.1 | 0.1 | | | 0.1 | | | 0.1 | | |
| Food and B. | 0.9 | 0.7 | 0.7 | 0.6 | 0.3 | 0.3 | | 0.2 | 0.4 | | | | | | |
| Informatio. | | | | | | | | | | | | | | | |
| Named Grou. | | | | | | | 0.7 | 0.4 | 0.2 | 0.1 | 0.1 | 0.1 | 0.2 | 0.2 | 0.2 |
| Health Car. | | | | | | | | | | 0.1 | | | 0.1 | 0.1 | 0.1 |
| Geographic. | | | | | | | | | | 0.1 | | | 0.3 | 0.1 | |

| POLICY | Top third count | | | | | | | | | Absolute Purity Difference | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Network | BERT | | | Mixed | | | Pure | | | Mixed | | | Pure | | |
| Coverage | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 |
| Anatomy. | 0.6 | 0.9 | 0.8 | 0.4 | 0.1 | | | | | | | | 0.1 | | |
| Organisms. | 0.9 | 0.7 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.9 | 0.1 | 0.1 | 0.1 | 0.2 | 0.1 | 0.1 |
| Diseases. | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.9 | | | 0.1 | 0.2 | 0.1 | 0.1 |
| Chemicals . | 0.3 | | | 0.9 | 0.8 | 0.4 | 1.0 | 0.9 | 0.6 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 |
| Analytical. | 1.0 | 0.9 | 0.8 | 1.0 | 0.9 | 0.8 | 0.8 | 0.8 | 0.6 | | | | 0.1 | | |
| Psychiatry. | | | | | | | | 0.1 | 0.2 | | | | 0.1 | | |
| Phenomena . | | | | | | | | | | | | | 0.1 | | |
| Natural Sc. | | | | | | | | | | 0.1 | | | 0.1 | | |
| Health Occ. | | | | | | | | | | | | | 0.1 | | |
| Social Sci. | 0.1 | | 0.1 | | | 0.1 | | 0.1 | 0.3 | | | | 0.1 | 0.1 | |
| Education. | 0.2 | 0.5 | 0.6 | | 0.3 | 0.6 | 0.2 | 0.2 | | | | | | | |
| Technology. | 0.6 | 0.6 | 0.3 | 0.4 | 0.5 | 0.4 | 0.7 | 0.5 | 0.4 | | | | 0.1 | 0.1 | 0.1 |
| Food and B. | 0.3 | 0.3 | 0.4 | 0.3 | 0.3 | 0.5 | 0.4 | 0.3 | 0.5 | 0.1 | 0.1 | | 0.1 | | |
| Informatio. | | | | | | | | | | | | | | | |
| Named Grou. | | | | | | | | | 0.3 | 0.1 | 0.1 | 0.1 | 0.2 | 0.1 | 0.1 |
| Health Car. | | | | | | | | | 0.1 | | | | 0.1 | 0.1 | |
| Geographic. | | | | | | | | | 0.1 | 0.2 | | | 0.6 | 0.6 | 0.6 |

| PATENTS | Top third count | | | | | | | | | Absolute Purity Difference | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Network | BERT | | | Mixed | | | Pure | | | Mixed | | | Pure | | |
| Coverage | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 |
| Anatomy. | 0.8 | 0.7 | 0.6 | 0.4 | 0.4 | 0.3 | | 0.1 | 0.1 | 0.1 | | | | | |
| Organisms. | 0.5 | 0.2 | 0.1 | 0.7 | 0.5 | 0.6 | 0.8 | 0.7 | 0.5 | 0.2 | 0.1 | 0.1 | 0.1 | | |
| Diseases. | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.9 | 1.0 | 1.0 | 0.1 | 0.1 | 0.1 | | | |
| Chemicals . | | | | | | | 0.6 | 0.6 | 0.7 | 0.3 | 0.2 | 0.1 | 0.1 | | |
| Analytical. | 0.1 | | | 0.1 | | | | | | 0.1 | | | | | |
| Psychiatry. | 0.2 | 0.8 | 0.9 | 0.1 | 0.6 | 0.7 | 0.2 | 0.1 | | 0.1 | | | | | |
| Phenomena . | | | | | | | 0.2 | 0.3 | 0.5 | 0.1 | | | | | |
| Natural Sc. | | | | | | | | | | 0.1 | | | | | 0.1 |
| Health Occ. | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| Social Sci. | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| Education. | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| Technology. | 0.1 | | 0.2 | 0.4 | 0.3 | 0.4 | | | | 0.2 | 0.1 | | | | |
| Food and B. | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| Informatio. | 0.2 | 0.3 | 0.2 | 0.3 | 0.2 | | 0.2 | 0.2 | 0.2 | | | | | | |
| Named Grou. | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| Health Car. | | | | | | | | | | 0.1 | | | | | |
| Geographic. | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |

| TWAUTHOR | Top third count | | | | | | | | | Absolute Purity Difference | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Network | BERT | | | Mixed | | | Pure | | | Mixed | | | Pure | | |
| Coverage | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 | .25 | .50 | .75 |
| Anatomy. | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.7 | 0.7 | 0.3 | | | | | | |
| Organisms. | 0.2 | 0.1 | 0.4 | 0.2 | 0.4 | 0.7 | 0.2 | 0.3 | 0.5 | | | | | | |
| Diseases. | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.9 | 0.6 | 0.1 | | | | | | |
| Chemicals . | | | | | | | | | | | | | | | |
| Analytical. | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 0.9 | 0.7 | | | | | | | |
| Psychiatry. | 0.8 | 0.7 | 0.6 | 0.6 | 0.7 | | | 0.1 | 0.2 | | | | | | |
| Phenomena . | | 0.2 | 0.5 | 0.1 | 0.2 | 0.1 | | | 0.5 | | | | | | |
| Natural Sc. | | | | | | | | 0.2 | 0.4 | | | | | | |
| Health Occ. | 0.1 | 0.1 | | 0.2 | 0.2 | 0.2 | 1.0 | 0.9 | 0.5 | | | | 0.1 | 0.1 | 0.1 |
| Social Sci. | | | | | | | 0.1 | 0.2 | 0.6 | | | | | | |
| Education. | . | . | . | . | . | . | . | . | . | . | . | . | . | . | . |
| Technology. | 0.3 | 0.5 | 0.1 | 0.3 | 0.5 | 0.4 | 0.1 | | | | | | | | |
| Food and B. | 0.6 | 0.3 | 0.3 | 0.6 | | 0.4 | 0.2 | 0.1 | 0.6 | | | | | | |
| Informatio. | | | | | | | 0.1 | 0.2 | 0.4 | | | | | | |
| Named Grou. | | | | | | 0.1 | 0.8 | 0.6 | 0.2 | | | | | | |
| Health Car. | | | | | | | 0.1 | 0.2 | 0.5 | | | | | | |
| Geographic. | | | | | | | | | | | | | | | |

Table 5.4: Summary of the results for each network. This table shows the absolute and relative Purity difference, but only the highest of the three Coverage values. All values are derived from Table 5.3. “M” and “P” indicate the Mixed and Pure networks, respectively. Light green and dark green indicate an absolute Purity difference of at least 0.2 and 0.5, respectively. One and two plus signs indicate a relative Purity difference of at least 0.2 and 0.5, respectively. The relative Purity difference is calculated from the top third count in Table 5.3. Dots mean that the topic category was not included in the experiment due to having too few topics per Size bin, as explained in the filtering process from Section 5.3.6.



| Source | Citat. | | Twconv | | Author | | Face. | | Policy | | Pat. | | Twau. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Network | M | P | M | P | M | P | M | P | M | P | M | P | M | P |
| Anatomy. | ▓ | ░ | ░ | | | | | | | | | | | |
| Organisms. | ▓ + | ░ ++ | ░ | | ++ | ░ ++ | ++ | ░ ++ | + | ░ + | ░ ++ | ++ | + | + |
| Diseases. | ▓ | ░ | ░ | | | | | | | ░ | | | | |
| Chemicals . | ▓ | ▓ + | ░ | + | | | | | ++ | ++ | ░ | ++ | | |
| Analytical. | ▓ | | ░ | | | | + | | | | | | | |
| Psychiatry. | ▓ | ░ | ░ | + | | + | | | | + | | | | |
| Phenomena . | ▓ | ░ | ░ | + | | | | | | | | ++ | | |
| Natural Sc. | ▓ | ░ | ░ | ++ | | | + | | | | | | | + |
| Health Occ. | ▓ | ░ | . | . | | ++ | ░ + | ░ ++ | | | . | . | + | ++ |
| Social Sci. | ▓ | ░ | ░ | ++ | | | | | | + | . | . | | ++ |
| Education. | . | . | . | . | . | . | . | . | | | . | . | . | . |
| Technology. | ▓ | ░ | ░ | | | | | | | | ░ + | | + | |
| Food and B. | ▓ | ░ | ▓ | | | | | | | | . | . | | + |
| Informatio. | ▓ | | ░ | | | | | | | | | | | + |
| Named Grou. | ▓ | ░ | ░ | ++ | | ++ | | ░ ++ | | ░ + | . | . | | ++ |
| Health Car. | ▓ | ░ | ░ | | | | | | | | | | | ++ |
| Geographic. | ▓ | ▓ | ▓ | + | ░ | ▓ ++ | | ░ | ░ | ▓ | . | . | | |

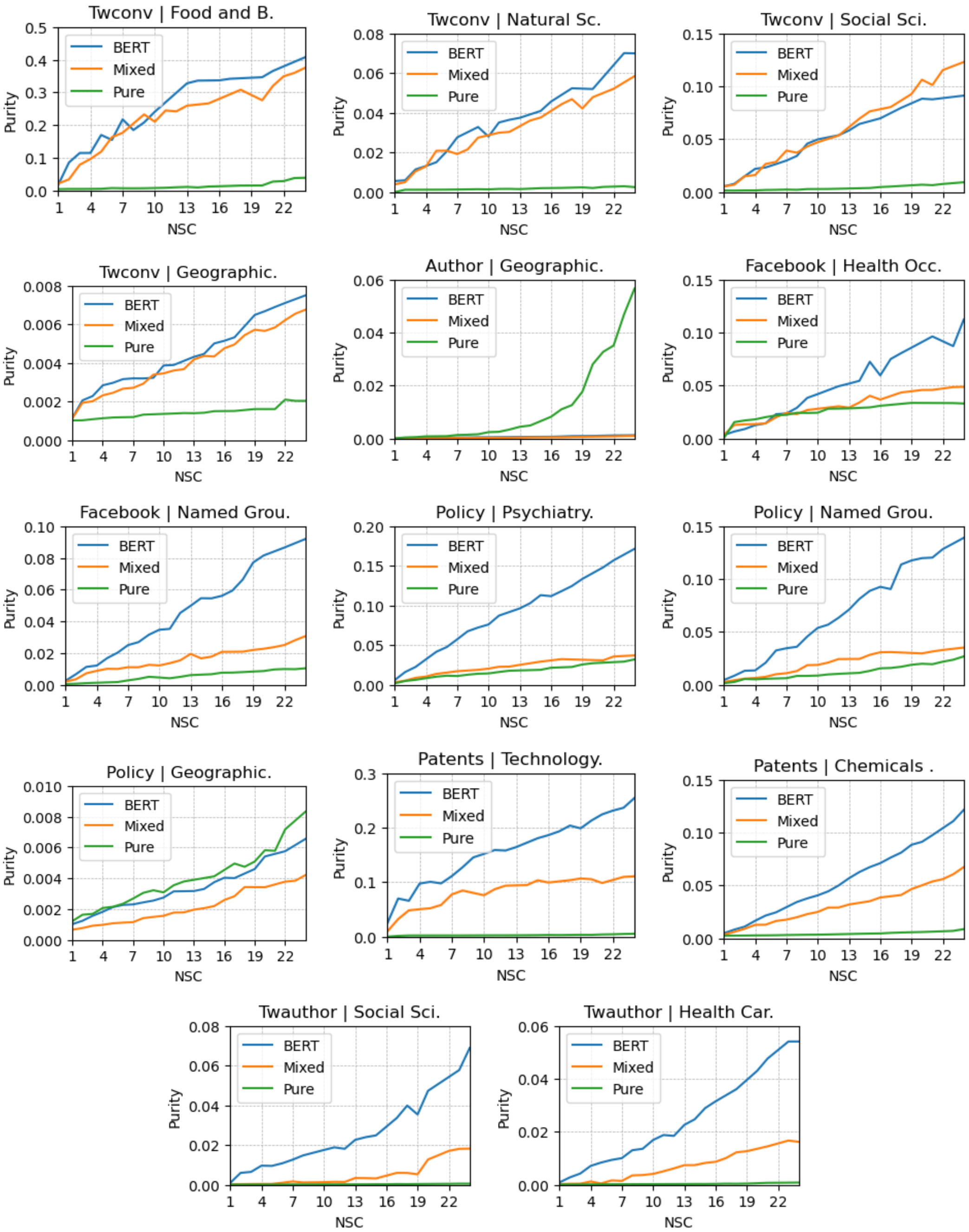


Figure 5.3: Examples of Purity of several topic categories for different networks. All profiles are for Size bin 161-320 and Coverage 0.50. To interpret these plots, it is important to keep in mind that each profile represents the average Purity and NSC across all topics in the topic category and Size bin, based on multiple clustering solutions.One way to interpret each curve is as if it were the Purity profile of a single, imaginary topic that combines all the topics in the category, including both the high- and low-performing ones. This topic would contain 240 documents (the average size of the bin), with each NSC value in the curve including 120 topic documents (due to Coverage 0.50). Purity values should not be compared across different sources, as some networks are substantially smaller, reducing clustering quality due to lack of information and making such comparisons unfair.

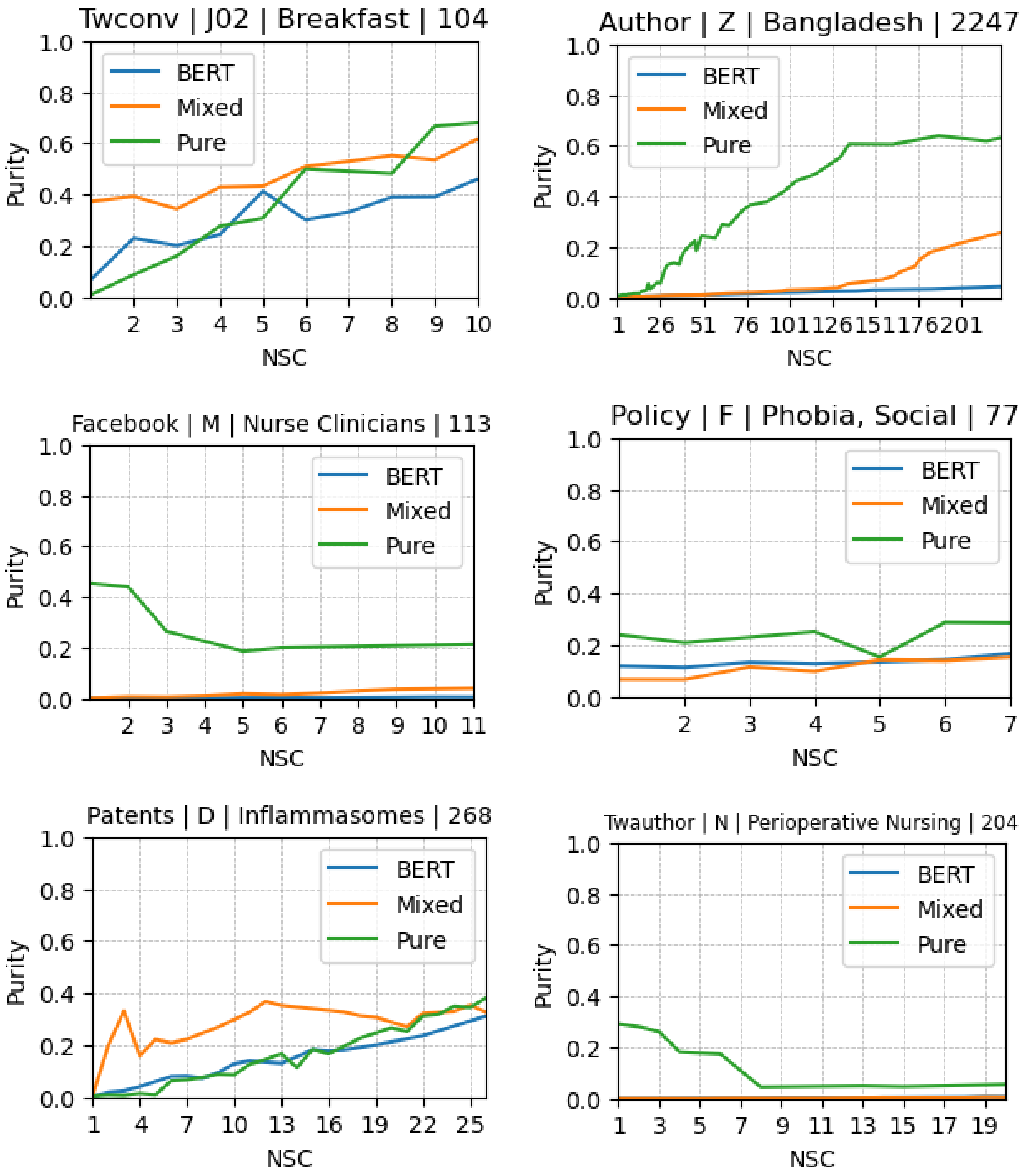


Figure 5.4: Examples of Purity profiles for individual topics across different networks. All Purity profiles are calculated for Coverage 0.50. The title of each plot indicates the external source, topic category, topic name and topic size.

Table 5.5: Best (non-citation) networks per topic category from Table 5.4. We selected the network(s) with the highest absolute difference, relative difference, or a combination of both, giving more weight to the absolute difference (i.e. in Table 5.4, dark green is preferred over two plus signs). The magnitude of the effect is shown as follows: **Zero stars**: Light green or one/two plus signs; **One star**: Light green and one plus symbol; **Two stars**: Light green with two plus signs, or dark green with zero/one plus signs; **Three stars**: Dark green with two plus signs.

| **Category** | **Best Networks** | **Magnitude** |
|---|---|---|
| Anatomy | mTwconv | |
| Organisms | mPatents, pFacebook, pAuthor | ** |
| Diseases | pPolicy, mTwconv | |
| Chemicals | mPatents, pPatents, mPolicy, pPolicy, mTwconv | |
| Analytical | mFacebook, mTwconv | |
| Psychiatry | pPolicy, mTwconv, pTwconv, pAuthor | |
| Phenomena | pPatents, mTwconv | |
| Natural Sc. | mTwconv, pTwconv | |
| Health Occ. | pFacebook | ** |
| Social Sci. | mTwconv, pTwconv, pTwauthor | |
| Education | - | |
| Technology | mPatents | * |
| Food and B. | mTwconv | ** |
| Informatio. | mTwconv, pTwauthor | |
| Named Grou. | pFacebook | ** |
| Health Car. | mTwconv, pTwauthor | |
| Geographic | pAuthor | *** |

### 5.4.1 Citations

As Table 5.4 shows, mCitation outperformed BERT and was the best-performing network overall. This aligns with prior findings in the literature, where networks that combine citations and text similarity tend to outperform either source alone [27]. pCitation also performed better than the other external sources, especially for *Chemicals and Drugs [D]*. However, in most topic categories it did not surpass BERT (i.e. absolute Purity difference $< 0.5$), which supports the use of BERT as a baseline in our analysis (with the exception of the topic category *Organisms [B]*).

The performance gap between BERT and pCitation is also interesting in light of our prior work [19], where we compared citation networks (using the same construction method) with text similarity networks based on the BM25 metric (a metric that matches and weights the words in common between documents). In that work, we found similar clustering effectiveness between the two. This suggests that BERT outperforms BM25, which is reasonable given that BERT is a more sophisticated method, although we did not test this comparison directly.

The fact that most networks outperform BERT for *Organisms [B]* may be due to BERT being a contextual embedding model, which means it represents words based on their surrounding context. Given that the context around different organism names is often very similar, BERT may struggle distinguishing between them. For this topic category, simpler term-frequency-based methods like BM25 might actually be more effective than contextual embeddings.

### 5.4.2 Twitter conversations

The mTwconv network had the best overall performance after the citation networks, achieving an absolute Purity difference of at least 0.2 in every topic category. We believe this is because Twitter conversations are more topically focused than the elements of other external sources. mTwconv performed best in the topic category *Food and Beverages [J02]*, likely due to the prevalence of nutrition-related discussions on Twitter.

Given this high performance, it is interesting that on the other hand, pTwconv did not achieve an absolute Purity difference of 0.2 or higher in any topic category. Also, the topic categories with the strongest improvements in mTwconv (*Food and Beverages [J02]* and *Geographicals [Z]*) are not the same as in pTwconv (which are *Natural Science Disciplines [H01]*, *Social Sciences [I01]* and *Named Groups [M]*). These differences between mTwconv and pTwconv suggest that mTwconv benefits significantly from the text similarity component. One likely reason is the sparse connectivity in pTwconv: On average, each external source element connects to only about two documents, compared to around twenty in pTwauthor. This low edge density may limit the quality of the clusters in pTwconv. The addition of the text similarity links in mTwconv may increase connectivity, allowing more coherent clusters.

The topic category profiles for *Food and Beverages [J02]* and *Geographicals [Z]* are slightly higher in mTwconv than in bTwconv (Figure 5.3), indicating that mTwconv is a competitive network. In contrast, the corresponding profiles in pTwconv are substantially lower.

### 5.4.3 Document authors

The pAuthor network performed best for the topic category *Geographicals [Z]*, although it showed poor results for most other categories. We believe this performance arises from the tendency of document authors to maintain stable interests over time about given geographical regions. In contrast, the mAuthor network did not produce interesting results. Figure 5.3 shows that *Geographicals [Z]* achieve a substantially higher profile in pAuthor than in bAuthor or mAuthor, making it very competitive. This is especially interesting given that, based on our prior work [19], the topic category *Geographicals [Z]* is the worst topic category for text similarity and citation networks by a substantial margin. While document authorship has been used in science mapping before, prior studies typically cluster authors rather than documents, with network edges representing co-authorship counts [101].

### 5.4.4 Facebook users

The pFacebook network performed well in the topic category *Named Groups [M]*, particularly for topics related to medical personnel (e.g. hospitalists), and it was the best-performing network for *Health Occupations [H02]*, especially in subtopics like medical specialties and nursing (e.g. neonatal nursing). This suggests that some Facebook users frequently share documents related to health advice, which makes sense because Facebook has a lot of support groups for people who suffer certain diseases where they share advice.

The profile of mFacebook for *Health Occupations [H02]* was about half that of bFacebook (Figure 5.3), so we believe mFacebook to be competitive for *Health Occupations [H02]*.

Interestingly, although pFacebook had a higher absolute Purity difference for *Named Groups [M]*, the topic category Purity profile for this category was actually lower than that of mFacebook (Figure 5.3). This suggests that a few specific topics (especially those related to medical personnel) performed very well in pFacebook, while the overall category performed better in mFacebook. In support of this, the highest performing topics within both *Named Groups [M]* and *Health Occupations [H02]* achieve much higher Purity in pFacebook than in bFacebook or mFacebook (see example in Figure 5.4).

These findings imply that if we had more finely defined topic categories focused exclusively on medical personnel, specialties, or nursing, both pFacebook and mFacebook would likely outperform bFacebook by a wider margin. This shows a limitation of the current topic category system and highlight the importance of examining interesting results in more detail, instead of taking them at face value.

### 5.4.5 Policy documents

The pPolicy network performed well in the topic categories *Named Groups [M]* and *Geographicals [Z]*, and was one of the few networks that showed improvement in *Psychiatry and Psychology [F]*, although the improvement there was small. We observed that topics with high Purity profiles within each category tended to share certain themes: In *Psychiatry and Psychology [F]*, the topics were often related to government (e.g. combat disorders) or societal issues (e.g. social phobia); in *Named Groups [M]*, they focused on medical professions and vulnerable groups (e.g. undocumented immigrants, persons with mental disabilities, minors); and in *Geographicals [Z]*, they were about American states and Global South countries (e.g. Colorado, Lebanon). In contrast, the mAuthor network did not produce interesting results.

These best performing topics in pPolicy seem to reflect the nature of policy documents. The first two categories focus on governmental and social matters, while the results for *Geographicals [Z]* likely reflect the American-centric coverage of the policy database, which overrepresents the Anglo-Saxon world [128].

The profiles for *Named Groups [M]* and *Psychiatry and Psychology [F]* in pPolicy are substantially lower than in bPolicy, while they are similar for *Geographicals [Z]* (Figure 5.3). This suggests that pPolicy is not a competitive network for these topic categories. Additionally, the mPolicy network shows lower Purity than both pPolicy and bPolicy, which is unusual among our results, suggesting that in this case, the external source and text similarity do not complement each other effectively.

### 5.4.6 Patent families

The mPatents network performed well in the topic categories *Chemicals and Drugs [D]*, particularly in topics related to biochemical elements (e.g. CD47 antigen), and *Technology, Industry, and Agriculture [J01]*, especially for topics about chemical components (e.g. dendrimers). This suggests that mPatents is effective for topics related to biotechnology, likely because these are closely tied to the types of inventions described in patents. In contrast, the pPatents network performed poorly in terms of absolute Purity difference, although it achieved the highest relative Purity difference for *Phenomena and Processes [G]*, likely also related to biotechnology. The reason why patents perform well for biotechnology might be due to the Biomedical focus of PubMed.

As shown in Figure 5.3, the profiles for *Chemicals and Drugs [D]* and *Technology, Industry, and Agriculture [J01]* in mPatents reach about half the Purity level of bPatents. We believe this is sufficient for mPatents to be considered competitive.

### 5.4.7 Twitter authors

The pTwauthor network was one of best for the topic categories *Social Sciences [I01]* and *Health Care [N]*, for the latter particularly in topics related to nursing (e.g. emergency nursing). This high clustering effectiveness is likely due to the fact that nursing is one of the most widely shared scientific topics on social media [59], which could be supported by some Twitter users sharing documents exclusively related to nursing. In contrast, the mTwauthor network did not produce interesting results.

Neither pTwauthor or mTwauthor had topic categories with absolute Purity difference higher than 0.2, and the pTwauthor profiles for *Social Sciences [I01]* and *Health Care [N]* were substantially lower than those in bTwauthor (Figure 5.3), suggesting that pTwauthor is not competitive.

Given the strong performance of mTwconv and the bad performance of pTwauthor and mTwauthor, this suggests that Twitter-based networks are more useful for science maps when they are built from conversations rather than users, despite the fact that user-based networks are more commonly used in the literature [45]. This difference may be due to the fact that individual users often tweet about multiple unrelated topics, while conversations tend to stay more focused on a specific theme. pTwauthor also perform much worse than pFacebook, which is the other network where users are the nodes. One possible reason is that Twitter has a high proportion of bot accounts that automatically share academic documents, at least compared to Facebook.

### 5.4.8 Twitter networks versus the other networks

We noticed that the Pure Twitter networks (pTwconv and pTwauthor) provide a very different perspective from the other sources. Excluding the topic category *Organisms [B]*, these are the networks with the highest number of topic categories with a high relative Purity difference, indicating that their best performing topic categories are very different from text similarity. Also, these are the networks that achieved the highest improvement for topic category *Natural Science Disciplines [H01]*, which is especially relevant because science map users often expect to see this category represented, but citation and text similarity science maps are not good at representing it [19].

We believe this distinctiveness reflects a deeper dichotomy in how science is organized. On one hand, Twitter (and to some extent Facebook) captures how laypeople perceive and talk about scientific topics. On the other hand, traditional sources reflect the structure of science as it emerges from practical use, such as through citations, patents, or authorship patterns. This contrast highlights the potential value of social media–based networks in revealing how society engages with and mentally organizes scientific knowledge.

### 5.4.9 Cases where Purity decreases at higher NSC

We noticed that for some topic Purity profiles, Purity decreased at higher NSC values, which is the opposite of what we expected. As we explained in Section 5.3.7.1, Purity tends to increase with higher NSC because smaller clusters allows a finer selection of clusters.

These decreasing trends were most common in pTwauthor and pFacebook. Upon inspection, the likely cause is the following (explained here in a technically imprecise way for ease of reading): In some topics, some selected clusters consist of documents that are only connected through one or a few Twitter or Facebook users, and these are the documents' only connections. When we run the clustering with a higher Resolution parameter, the clustering algorithm can no longer recreate these clusters because they become too large relative to the new Resolution constraints. Since the documents are equally connected, it becomes arbitrary which document is excluded to satisfy the new clustering conditions. If the excluded document belonged to the topic, the following happens: The smaller cluster is still selected for the topic evaluation because it likely still contain several topic documents, but now it provides less Purity due to the ratio of topic to non-topic documents. Meanwhile, the excluded topic document has no other connections, so it cannot be part of other clusters. These two effects decrease the overall Purity, even as NSC increases.

In summary, Purity may decrease at higher NSC in networks where many documents are linked to the same external source element and have no other connections. The fact that this pattern is observed in pTwauthor and pFacebook suggests that there are topics where several relevant documents are shared exclusively by a single social media user.

## 5.5 Discussion

In this section we will discuss the high level ideas, strengths and weaknesses of our work. One of our most important results is that the external sources tend to cluster some topic categories better than others, and that these topic categories are different between sources. This suggests that external sources provide complementary perspectives on how to group documents together, and that these perspectives capture meaningful dimensions of how knowledge is organized or perceived. These different perspectives are not only useful to create science maps, like in this paper, but they could potentially be applied in other areas to reveal how society perceives and engages with science. For example, the Twitter perspective is very different from the other networks, Facebook users share health science content, and document authors show consistent focus on specific geographical regions.

Also, even as the external sources tend to not outperform BERT in most topic categories, this was not the goal of the paper, and it is possible that an alternative method for constructing science maps could reach this goal.

A strength of our research is the clustering effectiveness evaluation method, which is a substantial improvement over the clustering effectiveness evaluation method we used in our prior work [19] because our new approach is much easier to interpret. In our previous work, we use two metrics evaluate effectiveness, Purity and the inverse clustering count, while now we simplify the evaluation by using only Purity. We also used to only be able to compare clustering effectiveness between clustering solutions with the same documents and similar cluster sizes, while now we can compare the clustering solutions of several Resolution values across networks with different documents. In the prior work we also did not have Purity profiles, which provide a very intuitive description of the quality of the topic clusters that a user would experience in a science map. However, the current method does miss certain nuances captured in our previous study. For example, we did not evaluate if some sources are better than others at different cluster sizes (our prior work and Xie and Waltman [177] found that citations are better than text for smaller clusters).

A limitation of our work is that we performed our experiments on clustering solutions that are less sophisticated than science maps used by researchers. For example, some science map methodologies have a minimum size for clusters, and clusters smaller than this size are merged with other clusters [164]. We did not do this, and as a consequence, when the nodes of a cluster are all equally connected by a few hub nodes in the network, reducing the size of the cluster by increasing the Resolution will turn random nodes of this cluster into singletons. This is a problem because, if this node is a topic document, then Purity would decrease at higher NSC, creating very confusing results for some topics that do not reflect the cluster effectiveness that would be observed in a science map. We observed this situation mostly in the Twitter users source, where some documents were shared by only one or two users. We did not attempt to prevent this situation because doing so would increase the complexity of our experimental design.

Another limitation of our research is that our Mixed networks combine a non-bipartite network (the BERT networks, which are non-bipartite because the links go from document to document) with a bipartite network (the Pure networks, bipartite because the links go from document to external source element). There are studies that use either of these types of networks for creating science maps, but there are no studies about combining them, which could have unintended effects in the map. The closest there is in the literature is the extended citation networks, where there are links from document to document and from document to non-core document, but not from non-core document to non-core document. Also, bipartite networks are not very common in science mapping, and it is more common to, instead of having the unit of co-occurrence in the network (in our case, the external source element), to represent the co-occurrence in the edge weight as a unipartite network [145]. The most common way of mapping unipartite and bipartite networks to each other is to project the bipartite network as unipartite [7], and the methods for projecting a bipartite network as a unipartite network are an ongoing topic of study [40, 118].

The method we used to combine the networks into the Mixed network is also relatively straightforward, and the only modification that we make is that the sum of edges weights in both networks must be the same. Chao and Tang [36] proposed a method to cluster networks with unipartite and bipartite structures, like our Mixed networks, but we decided to instead use the Leiden algorithm due to its preeminent position in the field of science mapping. We can imagine alternative modifications, for example trying mixing different proportions of the the external source and the text similarity edges, or normalizing all the edges that came out from a node so that they add up to the same value for all nodes. We did not normalize because normalization is used to control for different practices in reference list length across different academic fields, and since our dataset mostly contains biomedical fields we chose to avoid introducing additional complexity into our analyses. However, future research could explore how to create better Mixed networks for a given external source.

Another limitation is that we are comparing results created with different sets of documents, and using a subset of documents could hinder the formation of high quality clusters. We considered using the same set of documents for all sources. The first approach was to only use the documents present in all external sources, but this set of documents was very small. The second approach was to use all core documents, and let the disconnected clusters in the pure networks to form singleton clusters, but we saw that the quality of a topic was mostly influenced by how many of their documents had edges, instead of the extent that these edges connect documents from the same topic. In the end, we attempted to make the comparisons as fair as possible creating a text similarity network for each external network that also uses the same core documents. However, this does not address the fact that smaller networks have less information than bigger networks, which might decrease the quality of the clusters for both the text similarity and external network. For this reason, we avoid making strong statements based on the magnitude of Purity value (e.g. Purity 0.5 is good, Purity 0.005 is bad).

Another limitation is that the data sources that we used might not be available for researchers that use science maps. For instance, access to social media data such as Twitter has become increasingly restricted, limiting reproducibility or adoption by other researchers. We believe our results are still relevant because new sources of data can open up in the future, which can also be evaluated sing the same framework.

## 5.6 Conclusions

The topical bias of science maps limits their usefulness for topical analyses. In the current paper we have explored different data sources for creating academic documents networks that represent different document relations, with the purpose of finding sources that can change the topical bias of a science map. Our method of analysis was comparing the clustering effectiveness of different MeSH topic categories within a network and between networks, using a methodology that we refined from our prior work. We explored traditional science maps data sources (text similarity and citation links) and non-traditional data sources based on the co-occurrence of academic documents on another element (policy document, patent families, Facebook users, Twitter conversations, Twitter users, and document authors), which we referred to as external sources. Our comparisons were between networks that use either text similarity, external sources, or a mix of both.

We found that different external sources can be used to favor the emergence of different topics, and the following combinations had a particularly strong effect: Health for Facebook users, biotechnology for patent families, government and social issues for policy documents, food for Twitter conversations, nursing for Twitter users, and most strongly geographical entities for document authors. We also found that Twitter conversations work particularly well when combined with text similarity and that our text similarity metric (Sentence BERT) seems to perform better than the similarity metrics used in prior work (like BM25), except for topics related to organisms. Also, the favored topic categories are not affected by changing the percentage of the topic documents used in the evaluation, as shown by the similarity between the different Coverage values. Finally, the best topic categories in the Twitter networks were very different from the other networks, which means that Twitter (and potentially other similar social media platforms, like the new BlueSky or Mastodon) might provide different perspectives for the study of the organization of scientific knowledge, getting us closer to latent representations of how society perceives and interacts with science.

Our results show that external sources of academic document networks can be used to control topic bias, which opens up the possibility of creating science maps tailored for different needs. The most direct way of applying our discoveries is to create science maps biased toward different topics using these external sources. However, with the exception of document authors and their high clustering effectiveness for geographical entities, most external sources need to be used in combination with text similarity sources to achieve a high clustering effectiveness relative to traditional sources, and it is still an open question which is the best method for combining them into a single network. The clusters of external sources could also be used beyond science maps, for example to identify potential misuse of scientific publications (e.g. in misinformation strategies), or to identify societal connections or sensitivities that are not reflected in the academic world (e.g. connecting papers of diets and health concerns).

## 5.7 Data availability

The data and the code used to create the results is available at a Zenodo repository [13].

## 5.8 CRediT author statement

**Juan Pablo Bascur:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing.
**Rodrigo Costas:** Conceptualization, Writing – review & editing.
**Suzan Verberne:** Conceptualization, Methodology, Supervision, Writing – review & editing.

# Chapter 6

# Conclusion

In this dissertation we have attempted to partially fill the knowledge gap that exists in the literature on the performance of science maps for information retrieval. In the current chapter, we will answer the research questions that we presented in Chapter 1 and explore potential further research on this topic.

## 6.1 Answers to research questions

**Research question 1: How can science maps be designed to support information retrieval?**

We answered this question in Chapter 2 by implementing the tool SciMacro (Scientific Macroscope), which allows a user to navigate a science map of academic documents in a way that is conductive for information retrieval by using the principles of the Scatter-Gather method. From this research, we learned that there are no significant hindrances for implementing information retrieval in a science map, at least in the way we implement it. However, we found two minor challenges. The first is how to communicate relevant information using the bubble chart visualization of the science map, which we addressed by placing the related clusters together and minimizing white space. The second challenge is how to let the user control the granularity of the clusters, which we addressed by letting the user decide on the number of clusters they desire. Then, in the back end, we produced several clustering solutions with different resolutions until we found one that generated a good distribution of cluster sizes for this number of clusters (this is the slowest step and it has the greatest potential for improvement). After we had found this resolution, we merged the clusters until we got the number of clusters that the user desired.

**Research question 2: How effective are science maps for producing systematic reviews?**

We found in Chapter 3 that science maps are more effective than Boolean queries for about half of the evaluated systematic reviews, which is a good performance given the stringent conditions of the experiment (i.e., because the Boolean queries define the relevant documents, the baseline has perfect recall). This, plus our finding that the intersection between the sets of documents retrieved by the Boolean query and the ones retrieved by science maps is small, shows that one approach cannot replace the other, and ideally both should be used together for greatest effectiveness. We also found that science maps can correct for some shortcomings of the Boolean queries, like finding documents that the original authors missed. An interesting observation is that there was no topical difference between the set of systematic reviews where science maps performed better than the Boolean queries and the set of systematic reviews where they performed worse. This observation motivated research question 3.

**Research question 3: Do science maps represent some topics better than others?**

We found in Chapter 4 that some ontological categories of topics are systematically clustered better than others, in particular the ontological topic categories “Diseases” and “Organisms”, and that this happens in both citation and text similarity networks. Therefore, the answer to this research question is positive. For information retrieval tasks, this means that it is possible to know beforehand if a science map approach is likely to be helpful, which makes science maps a more reliable information retrieval tool. We were surprised that citation and text similarity networks perform well for the same topic categories because this suggests that the clusters of both maps would be about more or less the same topics. However, we also found differences between these networks. For higher granularity and Coverage (i.e. higher Coverage means higher recall), citation networks yield better results than text similarity networks, and vice versa. We believe this might be due to the simplicity of the text similarity metric that we used (i.e. it only measures shared words between documents and does not measure more subtle similarities like semantic similarity). It seems that creating good clusters at higher granularity and Coverage is more difficult than at lower, and so a more sophisticated text similarity metric might be needed.

**Research question 4: How can the representation of specific topics be improved in a science map?**

We answered this question in Chapter 5 by using different types of academic documents networks, based on data from different sources, to create science map clusters. This allowed us to influence which topic categories were the best clustered in a science map. Given that both text and citation networks yield similar results in terms of which topics are best clustered (as we found in response to research question 3), we used a text similarity network as a baseline (instead of a citation network or both networks). We compared the new networks with the baseline network to measure both the changes regarding the cluster quality of the topics and changes regarding which topics are best clustered in the new network. The biggest improvement in clustering effectiveness happened in topics related to geographical entities in the document authors network. The other noteworthy improvements were health topics in the Facebook users network, biotechnology topics in the patent families network, government and social topics in the policy documents network, food topics in the Twitter conversations network, and nursing topics in the Twitter users network. However, most of the topics that achieved the highest clustering effectiveness in their networks still achieved lower clustering effectiveness than in the text similarity networks, which defeats the purpose of improving the clustering effectiveness of the topic. A notable exception was the network that mixed text similarity with Twitter conversations. The topics obtained in this network had a clustering effectiveness comparable with text similarity, and even better for topics about food. Apart from this exception, we have not found a way to influence which topics are better represented in a science map without decreasing the quality of the clustering.

**Overarching research question: What is the effectiveness of science maps for information retrieval, and how can we enhance it?**

We studied science maps that are based on document clusters, using documents mostly from the biomedical field of science. These science maps have been shown to be effective for finding the relevant documents of systematic reviews, and to perform particularly well on topics that belong to the ontological topic categories “Diseases” and “Organisms”. The effectiveness of a science map can be enhanced by turning the map into an interactive visualization of the clusters, where the user can create a new visualization based on the documents in selected clusters and control the granularity of the map.

## 6.2 Further research

**Follow ups to our findings**

As discussed in the introduction, the research agenda set out in this dissertation is focused on evaluating and improving science maps for information retrieval. With regard to evaluation, we limited ourselves to systematic reviews and academic topics, but further research can also explore other information retrieval tasks, such as exploratory search tasks. With regard to improvement, we found that using different networks from different sources has the potential for influencing which topics are best represented, but only the network that mixed text similarity with Twitter conversations could achieve a performance that is as good as the performance of citation or text networks. We believe that a performance similar to the latter networks might be achieved by further refinement, for example by cleaning the data before creating the network (like removing bot users from Twitter), by creating the network with a different methodology (like normalizing the weights of the edges), or by mixing networks with a different criterion (like weighing one network more than the other). The issue of which ontological categories of topics are best represented in a science map has received only limited attention in the literature [76, 131], and future research in this area could provide new insights.

**Clustering**

A relevant topic that we did not research in this dissertation is the clustering algorithm [74] . The Leiden algorithm is the most popular one, but the MALBA algorithm [75] was created specifically to outperform the Leiden algorithm in field delimitation, and future research could use the methodology that we developed in Chapter 5 to compare them.

**Large Language Models**

Thanks to accelerating developments in large language models (LLMs), we believe that the text representation of documents will take a more prominent role in the creation of science maps. We can imagine that there could be a fine-tuned text embedding model for each of the ontological categories of topics that we analyzed (for example, there are over 6,000 pre-trained Sentence Transformer models available in the Hugging Face website [132]). Another area where these text processing methods can be used is in the cluster labeling, as shown by van Eck and Waltman [159], who labeled clusters by providing ChatGPT with their top 250 most cited documents. Additionally, entity recognition, which allows us to extract data directly from the documents, could improve science maps in unforeseen ways. Also, even though we did not compare it directly, our results in Chapter 5 strongly suggest that text similarity networks based on text embedding create better clusters than networks based on less advanced text processing methods.

Beyond text representation, LLMs are also relevant to science maps due to developments in retrieval-augmented generation (RAG), a method that retrieves documents to improve the quality of question answering of LLMs. The use of RAG for academic information retrieval is still an emerging field of study [22], but recent results show promise [9]. Also, Asai et al. [8] developed OPENSCHOLAR, a RAG tool specific for academic search. We believe that RAG does not replace science maps, but instead they complement each other, with science maps visualizing the RAG results and putting them in context. This search approach is already implemented in platforms such as Zeta Alpha [183].

**Granularity**

An important open issue in science mapping is the choice of granularity, understood as the level of detail of the map, usually corresponding to the size of clusters. There is no agreed-upon answer in the field, and accordingly, this dissertation addressed granularity in several different ways rather than fixing it to a single definition. In Chapters 2 and 3 it was controlled by a hypothetical user and by a user model, respectively. In Chapter 4 and 5 it was used to make fair comparisons, with the former centered on map granularity (size of clusters) and the latter focused on topic granularity (number of selected clusters). Other researchers have proposed different strategies: Sjögårde and Ahlgren [142] searched a granularity that would group the references of a review article into a single cluster, Held and Gläser [75] developed an algorithm to determine an adequate level based on network properties, and Ficozzi et al. [61] explored maximum granularity by representing each document individually, physicalized as a 100-square-meter floor mat. These diverse approaches show that granularity remains an open question, but also that it is central to making science maps useful for information retrieval.

**Prototyping**

We believe that the critical next step in research for science maps for information retrieval is the further development of prototypes. This would allow evaluating the performance of science maps with real users. This has the added benefit that, by showing concrete uses of science maps, it can bring additional interest to continue and support the research and sustainability of the software. We find this important because most of the proposals for academic information retrieval tools that we found in literature, even the ones we found promising, are currently unusable due to lack of maintenance. This could be achieved by collaborating with already existing academic information retrieval platforms, such as Web of Science, Scopus, Dimensions, Zeta Alpha, Semantic Scholar, Google Scholar, or OpenAlex. However, it is worth pointing out that evaluating the performance of science maps with real users is not a trivial task. Such evaluation of interactive information retrieval requires careful experimental design and the participation of field experts [94].

**Trends**

In this dissertation we have provided evidence and advice on how to make information retrieval with science maps a more viable option for academic users. Fortunately, since the start of our research, we have seen that bibliometrics enhanced information retrieval has gained popularity among researchers, and the open science movement is lobbying to make the metadata of academic documents openly available, which will make science maps more viable. We hope that our research will further strengthen these developments and will help support and popularize science maps for information retrieval.

# Bibliography

[1] Muhammad Kamran Abbasi and Ingo Frommholz. Cluster-based polyrepresentation as science modelling approach for information retrieval. *Scientometrics*, 102(3):2301–2322, 2015. doi: 10.1007/s11192-014-1478-1.

[2] Jimi Adams and Kate Vinita Fitch. Whose social capital?: Citation and co-citation patterns of a fragmented concept. *Socius*, 9:23780231231184766, 2023. doi: 10.1177/23780231231184766.

[3] Tanay Aggarwal, Angelo Salatino, Francesco Osborne, and Enrico Motta. Large language models for scholarly ontology generation: An extensive analysis in the engineering field. *arXiv*, 2024. doi: 10.48550/arXiv.2412.08258.

[4] Per Ahlgren, Yunwei Chen, Cristian Colliander, and Nees Jan van Eck. 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, 2020. doi: 10.1162/qss_a_00027.

[5] Altmetric.com. About us. `https://www.altmetric.com/about-us/`, 2024. Accessed: 2024-11-01.

[6] Miriam Arnold, Mascha Goldschmitt, and Thomas Rigotti. Dealing with information overload: a comprehensive review. *Frontiers in Psychology*, 14:1122200, 2023. doi: 10.3389/fpsyg.2023.1122200.

[7] Jesús Arroyo, Carey E. Priebe, and Vince Lyzinski. Graph matching between bipartite and unipartite networks: To collapse, or not to collapse, that is the question. *IEEE Transactions on Network Science and Engineering*, 8(4):3019–3033, 2021. doi: 10.1109/tnse.2021.3086508.

[8] Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Dan Weld, Doug Downey, Wen tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. Openscholar: Synthesizing scientific literature with retrieval-augmented lms, 2024. URL `https://arxiv.org/abs/2411.14199`.

[9] Ahmet Yasin Aytar, Kemal Kilic, and Kamer Kaya. A retrieval-augmented generation framework for academic literature navigation in data science, 2024. URL `https://arxiv.org/abs/2412.15404`.

[10] Leif Azzopardi, Kalervo Järvelin, Jaap Kamps, and Mark D. Smucker. Report on the SIGIR 2010 workshop on the simulation of interaction. *ACM SIGIR Forum*, 44(2):35–47, 2011. doi: 10.1145/1924475.1924484.

[11] Zhichao Ba and Zhentao Liang. A novel approach to measuring science-technology linkage: From the perspective of knowledge network coupling. *Journal of Informetrics*, 15(3):101167, 2021. doi: 10.1016/j.joi.2021.101167.

[12] Juan Pablo Bascur. Which topics are best represented by science maps? An analysis of clustering effectiveness for citation and text similarity networks (data). *Zenodo*, 2024. doi: 10.5281/zenodo.11181030.

[13] Juan Pablo Bascur. Use of diverse data sources to control which topics emerge in a science map. Supplementary material. *Zenodo*, 2024. doi: 10.5281/zenodo.14170721.

[14] Juan Pablo Bascur. Academic information retrieval using citation clusters: In-depth evaluation based on systematic reviews (data). *Zenodo*, 2025. doi: 10.5281/zenodo.6702251.

[15] Juan Pablo Bascur. Prototype of Scimacro. `https://github.com/jpbascur/Scimacro`, 2025. Accessed: 2025-04-11.

[16] Juan Pablo Bascur, Nees Jan van Eck, and Ludo Waltman. An interactive visual tool for scientific literature search: Proposal and algorithmic specification. In *BIR@ ECIR*, pages 76–87, 2019. URL `https://ceur-ws.org/Vol-2345/paper7.pdf`.

[17] Juan Pablo Bascur, Suzan Verberne, Nees Jan van Eck, and Ludo Waltman. Academic information retrieval using citation clusters: In-depth evaluation based on systematic reviews. *Scientometrics*, 128(5):2895–2921, 2023. doi: 10.1007/s11192-023-04681-x.

[18] Juan Pablo Bascur, Rodrigo Costas, and Suzan Verberne. Use of diverse data sources to control which topics emerge in a science map. *arXiv*, 2024. doi: 10.48550/arXiv.2412.07550.

[19] Juan Pablo Bascur, Suzan Verberne, Nees Jan van Eck, and Ludo Waltman. Which topics are best represented by science maps? An analysis of clustering effectiveness for citation and text similarity networks. *Scientometrics*, 130:1181–1199, 2025. doi: 10.1007/s11192-024-05218-6.

[20] Christopher W. Belter. Citation analysis as a literature search method for systematic reviews. *Journal of the Association for Information Science and Technology*, 67(11):2766–2777, 2016. doi: 10.1002/asi.23605.

[21] Christopher W. Belter. A relevance ranking method for citation-based search results. *Scientometrics*, 112(2):731–746, 2017. doi: 10.1007/s11192-017-2406-y.

[22] Ravi Varma Kumar Bevara, Brady D Lund, Nishith Reddy Mannuru, Sai Pranathi Karedla, Yara Mohammed, Sai Tulasi Kolapudi, and Aashrith Mannuru. Prospects of retrieval augmented generation (rag) for academic library search and retrieval. *Information Technology and Libraries*, 44(2), 2025.

[23] Nauman bin Ali and Binish Tanveer. A comparison of citation sources for reference and citation-based search in systematic literature reviews. *e-Informatica Software Engineering Journal*, 16(1):220106, 2022. doi: 10.37190/e-Inf220106.

[24] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[25] Katy Börner. Atlas of science: Visualizing what we know. *M Press*, 2010. doi: 10.5555/1995300.

[26] Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222, 2015. doi: 10.1002/asi.23329.

[27] Kevin W. Boyack and Richard Klavans. A comparison of large-scale science models based on textual, direct citation and hybrid relatedness. *Quantitative Science Studies*, 1(4):1570–1585, 2020. doi: 10.1162/qss_a_00085.

[28] Kevin W. Boyack, Catherine Smith, and Richard Klavans. A detailed open access model of the PubMed literature. *Scientific Data*, 7(1):408, 2020. doi: 10.1038/s41597-020-00749-y.

[29] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117, 1998. doi: 10.1016/S0169-7552(98)00110-X.

[30] Guillaume Cabanac, Muthu Kumar Chandrasekaran, Ingo Frommholz, Kokil Jaidka, Min-Yen Kan, Philipp Mayr, and Dietmar Wolfram. Report on the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL 2016). *SIGIR Forum*, 50(2):36–43, 2017. doi: 10.1145/3053408.3053417.

[31] Michel Callon, Jean-Pierre Courtial, William A. Turner, and Serge Bauin. From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22(2):191–235, 1983. doi: 10.1177/053901883022002003.

[32] Zeljko Carevic, Maria Lusky, Wilko van Hoek, and Philipp Mayr. Investigating exploratory search activities based on the stratagem level in digital libraries. *International Journal on Digital Libraries*, 19:231–251, 2018. doi: 10.1007/s00799-017-0226-6.

[33] David Carmel, Elad Yom-Tov, Adam Darlow, and Dan Pelleg. What makes a query difficult? In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*, page 390. ACM, 2006. doi: 10.1145/1148170.1148238.

[34] Chiara Carusi and Giuseppe Bianchi. Scientific community detection via bipartite scholar/journal graph co-clustering. *Journal of Informetrics*, 13(1):354–386, 2019. doi: 10.1016/j.joi.2019.01.004.

[35] Centre for Science and Technology Studies. Leiden ranking fields, 2023. URL `https://www.leidenranking.com/information/fields`. Accessed: 03-04-2023.

[36] Chang Chang and Chao Tang. Community detection for networks with unipartite and bipartite structure. *New Journal of Physics*, 16(9):093001, 2014. doi: 10.1088/1367-2630/16/9/093001.

[37] Chaomei Chen. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3):359–377, 2006. doi: 10.1002/asi.20317.

[38] Chaomei Chen. Science mapping: A systematic review of the literature. *Journal of Data and Information Science*, 2(2):1–40, 2017. doi: 10.1515/jdis-2017-0006.

[39] Chaomei Chen, Fidelia Ibekwe-SanJuan, and Jianhua Hou. The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. *Journal of the American Society for Information Science and Technology*, 61(7):1386–1409, 2010. doi: 10.1002/asi.21309.

[40] Giulio Cimini, Alessandro Carra, Luca Didomenicantonio, and Andrea Zaccaria. Meta-validation of bipartite network projections. *Communications Physics*, 5(1):76, 2022. doi: 10.1038/s42005-022-00856-9.

[41] Clarivate. Web of Science. `https://clarivate.com/products/web-of-science/`, 2018. Accessed: 2018-01-27.

[42] Manuel J. Cobo, Antonio G. López-Herrera, Enrique Herrera-Viedma, and Francisco Herrera. Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology*, 62(7):1382–1402, 2011. doi: 10.1002/asi.21525.

[43] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.207.

[44] Connected Papers. Connected papers: Visual tool for literature discovery. `https://www.connectedpapers.com/`, 2025. Accessed: 2025-04-11.

[45] Rodrigo Costas, Sarah de Rijcke, and Noortje Marres. "Heterogeneous couplings": Operationalizing network perspectives to study science-society interactions through social media metrics. *Journal of the Association for Information Science and Technology*, 72(5):595–610, 2021. doi: 10.1002/asi.24427.

[46] Sarah E. Cousins, Elizabeth Tempest, and David J. Feuer. Surgery for the resolution of symptoms in malignant bowel obstruction in advanced gynaecological and gastrointestinal cancer. *Cochrane Database of Systematic Reviews*, 2016. doi: 10.1002/14651858.CD002764.pub2.

[47] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 2006. URL `https://igraph.org`.

[48] Douglass R. Cutting, David R. Karger, Jan O. Pedersen, and John W. Tukey. Scatter/Gather: A cluster-based approach to browsing large document collections. *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 318 – 329, 1992. doi: 10.1145/133160.133214.

[49] Christiaan M. De Vries, Shlomo Geva, and Andrew Trotman. Document clustering evaluation: Divergence from a random baseline. *arXiv*, 2012. doi: 10.48550/arXiv.1208.5654.

[50] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.

[51] Digital Science. Dimensions. `https://www.dimensions.ai/`, 2018. Accessed: 2018-01-27.

[52] Ying Ding. Community detection: Topological vs. topical. *Journal of Informetrics*, 5(4): 498–514, 2011. doi: 10.1016/j.joi.2011.02.006.

[53] Pablo Dorta-González, Alejandro Rodríguez-Caro, and María Isabel Dorta-González. Societal and scientific impact of policy research: A large-scale empirical study of some explanatory factors using Altmetric and Overton. *Journal of Informetrics*, 18(3):101530, 2024. doi: 10.1016/j.joi.2024.101530.

[54] Ciriaco Andrea D'Angelo and Nees Jan Van Eck. Collecting large-scale publication data at the level of individual researchers: A practical proposal for author name disambiguation. *Scientometrics*, 123:883–907, 2020. doi: 10.1007/s11192-020-03410-y.

[55] David Ellis. Modeling the information-seeking patterns of academic researchers: A grounded theory approach. *The Library Quarterly*, 63(4):469–486, 1993. doi: 10.1086/602622.

[56] Elsevier. Scopus. `https://www.scopus.com/`, 2018. Accessed: 2018-01-27.

[57] Elsevier. Topic prominence in science—Scival. `https://www.elsevier.com/solutions/scival/features/topic-prominence-in-science`, 2023. Accessed: 2023-01-25.

[58] Martin J. Eppler. A comparison between concept maps, mind maps, conceptual diagrams, and visual metaphors as complementary tools for knowledge construction and sharing. *Information Visualization*, 5(3):202–210, 2006. doi: 10.1057/palgrave.ivs.9500131.

[59] Zhichao Fang, Rodrigo Costas, Wencan Tian, Xianwen Wang, and Paul Wouters. An extensive analysis of the presence of altmetric data for Web of Science publications across subject fields and research topics. *Scientometrics*, 124(3):2519–2549, 2020. doi: 10.1007/s11192-020-03564-9.

[60] Zhichao Fang, Jonathan Dudek, Ed Noyons, and Rodrigo Costas. Science cited in policy documents: Evidence from the Overton database. *arXiv*, 2024. doi: 10.48550/arXiv.2407.09854.

[61] Matilde Ficozzi, Mathieu Jacomy, Dario Rodighiero, Anne Beaulieu, and Anders Kristian Munk. Grounding ai map: The consequences of living with the trouble of an irreductionist map. *Design et abstractions, Revue Design Arts Medias*, 2025(6), 2025.

[62] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010. doi: 10.1016/j.physrep.2009.11.002.

[63] Ingo Frommholz, Philipp Mayr, Guillaume Cabanac, and Suzan Verberne. Bibliometric-enhanced information retrieval: 11th international BIR workshop. In *Advances in Information Retrieval*, pages 705–709. Springer International Publishing, 2021. doi: 10.1007/978-3-030-72240-1_85.

[64] Jochen Gläser. Opening the black box of expert validation of bibliometric maps. In *Lockdown Bibliometrics: Papers not submitted to the STI Conference 2020 in Aarhus*, pages 27–36, 2020.

[65] Audilio Gonzales-Aguilar and María Ramírez-Posada. Carot2: Búsqueda y visualización de la información. *Profesional de la información*, 21(1):105–112, 2012. doi: 10.3145/epi.2012.ene.14.

[66] Google. Google scholar. `https://scholar.google.com/`, 2018. Accessed: 2018-01-27.

[67] Christian Sengstock Hamed Abdelhaq and Michael Gertz. EvenTweet: Online localized event detection from Twitter. *Proceedings of the VLDB Endowment*, 6(12):1326–1329, 2013. doi: 10.14778/2536274.2536307.

[68] Robin Haunschild and Werner Marx. Discovering seminal works with marker papers. *Scientometrics*, 125(3):2955–2969, 2020. doi: 10.1007/s11192-020-03358-z.

[69] Robin Haunschild, Hermann Schier, Werner Marx, and Lutz Bornmann. 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, 2018. doi: 10.1016/j.joi.2018.03.004.

[70] Frank Havemann, Jochen Gläser, and Michael Heinz. Memetic search for overlapping topics based on a local evaluation of link communities. *Scientometrics*, 111:1089–1118, 2017. doi: 10.1007/s11192-017-2302-5.

[71] Jiangen He, Qing Ping, Wen Lou, and Chaomei Chen. PaperPoles: Facilitating adaptive visual exploration of scientific publications by citation links. *Journal of the Association for Information Science and Technology*, 70(8):843–857, 2019. doi: 10.1002/asi.24171.

[72] Jiyin He, Edgar Meij, and Maarten de Rijke. Result diversification based on query-specific cluster ranking. *Journal of the American Society for Information Science and Technology*, 62 (3):550–571, 2011. doi: 10.1002/asi.21468.

[73] Marti A. Hearst and Jan O. Pedersen. Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 1996. doi: 10.1145/243199.243216.

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

[75] Matthias Held and Jochen Gläser. Exploring publication networks with a local cohesion-maximizing algorithm. *Quantitative Science Studies*, 5(3):681–703, 2024. doi: 10.1162/qss_a_00314.

[76] Matthias Held and Theresa Velden. How to interpret algorithmically constructed topical structures of scientific fields? A case study of citation-based mappings of the research specialty of invasion biology. *Quantitative Science Studies*, 3(3):651–671, 2022. doi: 10.1162/qss_a_00194.

[77] Matthias Held, Grit Laudel, and Jochen Gläser. Topic reconstruction from networks of papers may not be possible if only one algorithm is applied to only one data model. In *Lockdown Bibliometrics: Papers not submitted to the STI Conference 2020 in Aarhus*, pages 18–26, 2020.

[78] Matthias Held, Grit Laudel, and Jochen Gläser. Challenges to the validity of topic reconstruction. *Scientometrics*, 126:4511–4536, 2021. doi: 10.1007/s11192-021-03920-3.

[79] Bradley M. Hemminger, Dongrong Lu, K. T. L. Vaughan, and Susan J. Adams. Information seeking behavior of academic scientists. *Journal of the American Society for Information Science and Technology*, 58(14):2205–2225, 2007. doi: 10.1002/asi.20686.

[80] César A Hidalgo. Disconnected, fragmented, or united? A trans-disciplinary review of network science. *Applied Network Science*, 1:1–19, 2016. doi: 10.1007/s41109-016-0010-3.

[81] Katja Hofmann, Shimon Whiteson, and Maarten de Rijke. Balancing exploration and exploitation in listwise and pairwise online learning to rank for information retrieval. *Information Retrieval*, 16(1):63–90, 2013. doi: 10.1007/s10791-012-9197-9.

[82] Thomas E. Ferrin Holly J. Atkinson, John H. Morris and Patricia C. Babbitt. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. *PLoS ONE*, 4(2):e4345, 2009. doi: 10.1371/journal.pone.0004345.

[83] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength natural language processing in Python. *Zenodo*, 2020. doi: 10.5281/zenodo.1212303.

[84] Daniel W. Hook, Simon J. Porter, and Christian Herzog. Dimensions: Building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, 3:23, 2018. doi: 10.3389/frma.2018.00023.

[85] Tanya Horsley, Owen Dingwall, and Margaret Sampson. Checking reference lists to find additional studies for systematic reviews. *Cochrane Database of Systematic Reviews*, 2011. doi: 10.1002/14651858.MR000026.pub2.

[86] Darko Hric, Richard K. Darst, and Santo Fortunato. Community detection in networks: Structural communities versus ground truth. *Physical Review E*, 90(6):062805, 2014. doi: 10.1103/PhysRevE.90.062805.

[87] Inciteful. Inciteful: Interactive academic literature discovery. `https://inciteful.xyz/`, 2025. Accessed: 2025-04-11.

[88] Iris AI AS. Iris.ai: Ai-powered research assistant. `https://iris.ai/`, 2025. Accessed: 2025-04-11.

[89] A. Cecile J. W. Janssens and Marta Gwinn. Novel citation-based search method for scientific literature: Application to meta-analyses. *BMC medical research methodology*, 15:1–11, 2015. doi: 10.1186/s12874-015-0077-z.

[90] A. Cecile J. W. Janssens, Marta Gwinn, John E. Brockman, Kenneth Powell, and Melody Goodman. Novel citation-based search method for scientific literature: A validation study. *BMC Medical Research Methodology*, 20(1):1–11, 2020. doi: 10.1186/s12874-020-0907-5.

[91] Frizo Janssens, Wolfgang Glänzel, and Bart De Moor. A hybrid mapping of information science. *Scientometrics*, 75(3):607–631, 2008. doi: 10.1007/s11192-007-2002-7.

[92] Nick Jardine and Cornelis Joost van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information Storage and Retrieval*, 7(5):217–240, 1971. doi: 10.1016/0020-0271(71)90051-9.

[93] Byeongwoo Kang and Gianluca Tarasconi. PATSTAT revisited: Suggestions for better usage. *World patent information*, 46:56–63, 2016. doi: 10.1016/j.wpi.2016.06.001.

[94] Diane Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1–2):1–224, 2009. doi: 10.1561/1500000012.

[95] Max Kemman, Martijn Kleppe, and Stef Scagliola. Just Google it. Digital research practices of humanities scholars. *Studies in the Digital Humanities*, 2014. URL `http://hdl.handle.net/1765/50779`.

[96] Maxwell Mirton Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14(1):10–25, 1963. doi: 10.1002/asi.5090140103.

[97] Richard Klavans and Kevin W. Boyack. Quantitative evaluation of large maps of science. *Scientometrics*, 68(3):475–499, 2006. doi: 10.1007/s11192-006-0125-x.

[98] Richard Klavans and Kevin W. Boyack. 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, 2017. doi: 10.1002/asi.23734.

[99] Ben Slater Krishnan Chandra and Mike Ma. Research rabbit. `https://www.researchrabbit.ai/`, 2025. Accessed: 2025-04-11.

[100] Carol C. Kuhlthau. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5):361–371, 1991.

[101] Sameer Kumar. Co-authorship networks: A review of the literature. *Aslib Journal of Information Management*, 67(1):55–73, 2015. doi: 10.1108/ajim-09-2014-0116.

[102] Kuei-Kuei Lai and Shiao-Jun Wu. Using the patent co-citation approach to establish a new patent classification system. *Information processing & management*, 41(2):313–330, 2005. doi: 10.1016/j.ipm.2003.11.004.

[103] Kiwon Lee and Suchul Lee. Knowledge structure of the application of high-performance computing: A co-word analysis. *Sustainability*, 13(20):11249, 2021. doi: 10.3390/su132011249.

[104] Loet Leydesdorff. Theories of citation? *Scientometrics*, 43:5–25, 1998. doi: 10.1007/BF02458391.

[105] Yujie Liang, Qian Li, and Tianrui Qian. Finding relevant papers based on citation relations. In *Web-Age Information Management*, volume 6897 of *Lecture Notes in Computer Science*, pages 403–414. Springer Berlin Heidelberg, 2011. doi: 10.1007/978-3-642-23535-1_35.

[106] Suzanne K. Linder, Geetanjali R. Kamath, Gregory F. Pratt, Smita S. Saraykar, and Robert J. Volk. Citation searches are more sensitive than keyword searches to identify studies using specific measurement instruments. *Journal of clinical epidemiology*, 68(4):412–417, 2015. doi: 10.1016/j.jclinepi.2014.10.008.

[107] Litmaps. Litmaps. `https://www.litmaps.com/`, 2025. Accessed: 2025-04-11.

[108] Ziyang Liu, Sivaramakrishnan Natarajan, and Yi Chen. Query expansion based on clustered results. *Proceedings of the VLDB Endowment*, 4(6):350–361, 2011. doi: 10.14778/1978665.1978667.

[109] Maria Angeles Lopez-Olivo, Matxalen Amezaga Urruela, Lynda McGahan, Eduardo N. Pollono, and Maria E. Suarez-Almazor. Rituximab for rheumatoid arthritis. *Cochrane Database of Systematic Reviews*, 2015. doi: 10.1002/14651858.CD007356.pub2.

[110] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[111] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. Scoring, term weighting and the vector space model. In *Introduction to Information Retrieval*, chapter 6. Cambridge University Press, 2008. URL `https://nlp.stanford.edu/IR-book/html/htmledition/scoring-term-weighting-and-the-vector-space-model-1.html`.

[112] Philipp Mayr and Andrea Scharnhorst. Scientometrics and information retrieval: Weak-links revitalized. *Scientometrics*, 102(3):2193–2199, 2015. doi: 10.1007/s11192-014-1484-3.

[113] Martin Meyer. Does science push technology? Patents citing scientific literature. *Research policy*, 29(3):409–434, 2000. doi: 10.1016/S0048-7333(99)00040-2.

[114] Jose A. Moral-Munoz, Antonio G. López-Herrera, Enrique Herrera-Viedma, and Manuel J. Cobo. Science mapping analysis software tools: A review. *Springer Handbook of Science and Technology Indicators*, pages 159–185, 2019. doi: 10.1007/978-3-030-02511-3_7.

[115] Erwan Moreau. Literature-based discovery: Addressing the issue of the subpar evaluation methodology. *Bioinformatics*, 39(2):btad090, 2023. doi: 10.1093/bioinformatics/btad090.

[116] Peter Mutschke and Philipp Mayr. Science models for search: A study on combining scholarly information retrieval and scientometrics. *Scientometrics*, 102:2323–2345, 2015. doi: 10.1007/s11192-014-1485-2.

[117] National Institutes of Health. Medical subject headings. Available at `https://www.nlm.nih.gov/mesh/meshhome.html`, 2024. Accessed: 2024-11-01.

[118] Zachary Neal. The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. *Social Networks*, 39:84–97, 2014. doi: 10.1016/j.socnet.2014.06.001.

[119] Rommert Dekker Nees Jan van Eck, Ludo Waltman and Jan van den Berg. A comparison of two techniques for bibliometric mapping: Multidimensional scaling and VOS. *Journal of the American Society for Information Science and Technology*, 61(12):2405–2416, 2010. doi: 10.1002/asi.21421.

[120] Mark E.J. Newman. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the national academy of sciences*, 101(suppl_1):5200–5205, 2004. doi: 10.1073/pnas.0307545100.

[121] Mark E.J. Newman and Aaron Clauset. Structure and inference in annotated networks. *Nature Communications*, 7(1):11863, 2016. doi: 10.1038/ncomms11863.

[122] Open Knowledge Maps. Open knowledge maps: A visual interface to the world's scientific knowledge. `https://openknowledgemaps.org`, 2019. Accessed: 2025-04-16.

[123] OpenAlex. OpenAlex topic classification. `https://github.com/ourresearch/openalex-topic-classification`, 2024. Accessed: 20-04-2025.

[124] Francisco M. Ortuño, Ignacio Rojas, Miguel A. Andrade-Navarro, and Jean-François Fontaine. Using cited references to improve the retrieval of related biomedical documents. *BMC Bioinformatics*, 14(1):113, 2013. doi: 10.1186/1471-2105-14-113.

[125] Leto Peel, Daniel B Larremore, and Aaron Clauset. The ground truth about metadata and community detection in networks. *Science Advances*, 3(5):e1602548, 2017. doi: 10.1126/sciadv.1602548.

[126] Frank Peinemann, Carmen Bartel, Ulrich Grouven, and Frank Berthold. Retinoic acid post consolidation therapy for high-risk neuroblastoma. *Cochrane Database of Systematic Reviews*, 2013. doi: 10.1002/14651858.CD010685.

[127] Eugenio Petrovich. Science mapping and science maps. *Knowledge Organization*, 48(7-8): 535–562, 2022. doi: 10.5771/0943-7444-2021-7-8-535.

[128] Henrique Pinheiro, Etienne Vignola-Gagné, and David Campbell. A large-scale validation of the relationship between cross-disciplinary research and its uptake in policy-related documents, using the novel Overton altmetrics database. *Quantitative Science Studies*, 2(2):616–642, 2021. doi: 10.1162/qss_a_00137.

[129] Peter Pirolli, Patricia Schank, Marti Hearst, and Christopher Diehl. Scatter/Gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '96): Common Ground*, pages 213–220. ACM, 1996. doi: 10.1145/238386.238489.

[130] Ian Potter. Introducing citation topics in incites. `https://clarivate.com/blog/introducing-citation-topics/`, 2020. Accessed: 2025-04-16.

[131] Ismael Rafols. Towards multiple ontologies in science mapping. A tribute to Loet Leydesdorff. *SocArXiv*, 2025. doi: 10.31235/osf.io/47umh_v2.

[132] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410.

[133] Stephen Robertson and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009. doi: 10.1561/1500000019.

[134] Karen A. Robinson, Adam G. Dunn, Guy Tsafnat, and Paul Glasziou. Citation networks of related trials are often disconnected: Implications for bidirectional citation searches. *Journal of Clinical Epidemiology*, 67(7):793–799, 2014. doi: 10.1016/j.jclinepi.2013.11.015.

[135] Giulio Rossetti, Luca Pappalardo, and Salvatore Rinzivillo. A novel approach to evaluate community detection algorithms on ground truth. In *Complex Networks VII: Proceedings of the 7th Workshop on Complex Networks CompleNet 2016*, pages 133–144. Springer, 2016. doi: 10.1007/978-3-319-30569-1_10.

[136] Tony Russell-Rose, Jon Chamberlain, and Leif Azzopardi. Information retrieval in the workplace: A comparison of professional search practices. *Information Processing & Management*, 54(6):1042–1057, 2018. doi: 10.1016/j.ipm.2018.07.003.

[137] SAS Institute. Clustering methods. Available at `https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_cluster_sect012.htm`, 2009. Accessed: 2024-11-01.

[138] Eric Sayers. E-utilities quick start. `https://www.ncbi.nlm.nih.gov/books/NBK25500/`, 2008. Accessed: 2025-04-16.

[139] Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Leif Azzopardi, and Shlomo Geva. A test collection for evaluating retrieval of studies for inclusion in systematic reviews. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1237–1240. Association for Computing Machinery, 2017. doi: 10.1145/3077136.3080707.

[140] Harrisen Scells, Daniel Locke, and Guido Zuccon. An information retrieval experiment framework for domain specific applications. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1281–1284. ACM, 2018. doi: 10.1145/3209978.3210167.

[141] Cara Seitz, Marion Schmidt, Nathalie Schwichtenberg, and Theresa Velden. A case study of the epistemic function of citations—implications for citation-based science mapping. In *Proceedings of the 18th International Conference of the International Society for Scientometrics and Informetrics (ISSI)*, pages 1027–1032, 2021.

[142] Peter Sjögårde and Per Ahlgren. Granularity of algorithmically constructed publication-level classifications of research publications: Identification of topics. *Journal of Informetrics*, 12(1): 133–152, 2018. doi: 10.1016/j.joi.2017.12.006.

[143] Peter Sjögårde and Per Ahlgren. Granularity of algorithmically constructed publication-level classifications of research publications: Identification of specialties. *Quantitative Science Studies*, 1(1):207–238, 2020. doi: 10.1162/qss_a_00004.

[144] Peter Sjögårde, Per Ahlgren, and Ludo Waltman. Algorithmic labeling in hierarchical classifications of publications: Evaluation of bibliographic fields and term weighting approaches. *Journal of the Association for Information Science and Technology*, 72(7):853–869, 2021. doi: 10.1002/asi.24452.

[145] Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the Association for Information Science and Technology*, 24(4): 265–269, 1973. doi: 10.1002/asi.4630240406.

[146] Henry Small and Eugene Garfield. The geography of science: Disciplinary and national mappings. *Journal of Information Science*, 11(4):147–159, 1985. doi: 10.1177/016555158501100402.

[147] Jerzy Stefanowski and Dawid Weiss. Carrot2 and language properties in web search results clustering. In *Advances in Web Intelligence*, volume 2663 of *Lecture Notes in Computer Science*, pages 240–249. Springer Berlin Heidelberg, 2003. doi: 10.1007/3-540-44831-4_25.

[148] Lovro Šubelj, Nees Jan Van Eck, and Ludo Waltman. Clustering scientific publications based on citation relations: A systematic comparison of different methods. *PloS one*, 11(4):e0154404, 2016. doi: 10.1371/journal.pone.0154404.

[149] Richard Sullivan, Seth Eckhouse, and Grant Lewison. Using bibliometrics to inform cancer research policy and spending. *Monitoring financial flows for health research*, pages 67–78, 2007.

[150] Martin Szomszor and Euan Adie. Overton: A bibliometric database of policy document citations. *Quantitative Science Studies*, 3(3):624–650, 2022. doi: 10.1162/qss_a_00204.

[151] Bart Thijs. Science mapping and the identification of topics: Theoretical and methodological considerations. *Springer Handbook of Science and Technology Indicators*, pages 213–233, 2019. doi: 10.1007/978-3-030-02511-3_9.

[152] Anastasios Tombros, Rita Villa, and Cornelis Joost van Rijsbergen. The effectiveness of query-specific hierarchic clustering in information retrieval. *Information Processing & Management*, 38(4):559–582, 2002. doi: 10.1016/S0306-4573(01)00048-6.

[153] Vincent A. Traag, Ludo Waltman, and Nees Jan Van Eck. From Louvain to Leiden: Guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019. doi: 10.1038/s41598-019-41695-z.

[154] Nees Jan van Eck. Methodological advances in bibliometric mapping of science. *Erasmus Research Institute of Management*, 2011. URL `hdl.handle.net/1765/26509`.

[155] Nees Jan van Eck. Examples – VOSviewer Online Docs, 2022. URL `https://app.vosviewer.com/docs/examples/`. Accessed: 2025-04-11.

[156] Nees Jan van Eck and Ludo Waltman. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538, 2010. doi: 10.1007/s11192-009-0146-3.

[157] Nees Jan van Eck and Ludo Waltman. CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, 8(4):802–823, 2014. doi: 10.1016/j.joi.2014.07.006.

[158] Nees Jan van Eck and Ludo Waltman. Citation-based clustering of publications using CitNetExplorer and VOSviewer. *Scientometrics*, 111(2):1053–1070, 2017. doi: 10.1007/s11192-017-2300-7.

[159] Nees Jan van Eck and Ludo Waltman. An open approach for classifying research publications. *Leiden Madtrics*, 2024. doi: 10.59350/qc0px-76778.

[160] Cornelis Joost van Rijsbergen. Information retrieval: Theory and practice. In *Proceedings of the joint IBM/University of Newcastle Upon Tyne seminar on Data Base Systems*, pages 1–14, 1979.

[161] Cornelis Joost van Rijsbergen and W. Bruce Croft. Document clustering: An evaluation of some experiments with the cranfield 1400 collection. *Information Processing & Management*, 11(5–7):171–182, 1975. doi: 10.1016/0306-4573(75)90006-0.

[162] Theresa Velden, Kevin W. Boyack, Jochen Gläser, Rob Koopman, Andrea Scharnhorst, and Shenghui Wang. Comparison of topic extraction approaches and their results. *Scientometrics*, 111(2):1169–1221, 2017. doi: 10.1007/s11192-017-2306-1.

[163] Barnabas James Walker. Citationgecko. *Zenodo*, 2022. doi: 10.5281/zenodo.7068284. Accessed: 2025-04-16.

[164] Ludo Waltman and Nees Jan van Eck. 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, 2012. doi: 10.1002/asi.22748.

[165] Ludo Waltman, Kevin W. Boyack, Giovanni Colavizza, and Nees Jan van Eck. A principled methodology for comparing relatedness measures for clustering publications. *Quantitative Science Studies*, 1(2):691–713, 2020. doi: 10.1162/qss_a_00035.

[166] Feifei Wang, Chenran Jia, Xiaohan Wang, Junwan Liu, Shuo Xu, Yang Liu, and Chenyuyan Yang. Exploring all-author tripartite citation networks: A case study of gene editing. *Journal of Informetrics*, 13(3):856–873, 2019. doi: 10.1016/j.joi.2019.08.002.

[167] Qi Wang, Bentao Zou, Jialin Jin, and Yuefen Wang. Studying the linkage patterns and incremental evolution of domain knowledge structure: A perspective of structure deconstruction. *Scientometrics*, 129:4249–4274, 2024. doi: 10.1007/s11192-024-05059-3.

[168] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 77:34–49, 2018. doi: 10.1016/j.jbi.2017.11.011.

[169] Florian Wätzold, Bartosz Popiela, and Jonas Mayer. Methodology for AI-based search strategy of scientific papers: Exemplary search for hybrid and battery electric vehicles in the Semantic Scholar database. *Publications*, 12(4):49, 2024. doi: 10.3390/publications12040049.

[170] Michael E. Weinblatt, Roy Fleischmann, Tom W. J. Huizinga, Paul Emery, Janet Pope, Elisa M. Massarotti, Ronald F. van Vollenhoven, Joerg Wollenhaupt, Clifton O. Bingham, Brian Duncan, Namit Goel, Owen R. Davies, and Maxime Dougados. Efficacy and safety of certolizumab pegol in a broad population of patients with active rheumatoid arthritis: Results from the REALISTIC phase IIIb study. *Rheumatology*, 51(12):2204–2214, 2012. doi: 10.1093/rheumatology/kes150.

[171] Peter Willett. Recent trends in hierarchic document clustering: A critical review. *Information Processing & Management*, 24(5):577–597, 1988. doi: 10.1016/0306-4573(88)90027-1.

[172] Kate Williams. What counts: Making sense of metrics of research value. *Science and Public Policy*, 49(3):518–531, 2022. doi: 10.1093/scipol/scac004.

[173] Dietmar Wolfram. The symbiotic relationship between information retrieval and informetrics. *Scientometrics*, 102(3):2201–2214, 2015. doi: 10.1007/s11192-014-1479-0.

[174] Seokkyun Woo and John P. Walsh. On the shoulders of fallen giants: What do references to retracted research tell us about citation behaviors? *Quantitative Science Studies*, 5(1):1–30, 2024. doi: 10.1162/qss_a_00303.

[175] Katie Wright, Su Golder, and Rocio Rodriguez-Lopez. Citation searching: A systematic review case study of multiple risk behaviour interventions. *BMC Medical Research Methodology*, 14 (1):73, 2014. doi: 10.1186/1471-2288-14-73.

[176] Weili Wu, Hui Xiong, and Shashi Shekhar. *Clustering and information retrieval*. Springer Science & Business Media, 2003. doi: 10.1007/978-1-4613-0227-8.

[177] Qianqian Xie and Ludo Waltman. A comparison of citation-based clustering and topic modeling for science mapping. *Scientometrics*, 2025. doi: 10.1007/s11192-025-05324-z.

[178] Shuo Xu, Junwan Liu, Dongsheng Zhai, Xin An, Zheng Wang, and Hongshen Pang. Overlapping thematic structures extraction with mixed-membership stochastic blockmodel. *Scientometrics*, 117:61–84, 2018. doi: 10.1007/s11192-018-2841-4.

[179] Puyu Yang and Giovanni Colavizza. A map of science in Wikipedia. In *Companion Proceedings of the Web Conference 2022*, pages 1289–1300, 2022. doi: 10.1145/3487553.3524925.

[180] Zhiguo Yu. Understanding PubMed search results using topic models and interactive information visualization. *The University of Texas School of Biomedical Informatics at Houston*, 2017. URL `https://digitalcommons.library.tmc.edu/uthshis_dissertations/42/`.

[181] Ming Yuan, Justin Zobel, and Jimmy Lin. Measurement of clustering effectiveness for document collections. *Information Retrieval Journal*, 25:239–268, 2022. doi: 10.1007/s10791-021-09401-8.

[182] Jinhyuk Yun, Sejung Ahn, and June Young Lee. Return to basics: Clustering of scientific literature using structural information. *Journal of Informetrics*, 14(4):101099, 2020. doi: 10.1016/j.joi.2020.101099.

[183] Zeta Alpha. Zeta alpha. `https://www.zeta-alpha.com`, 2025. Accessed: 2025-09-29.

[184] Michel Zitt. Meso-level retrieval: IR-bibliometrics interplay and hybrid citation-words methods in scientific fields delineation. *Scientometrics*, 102(3):2223–2245, 2015. doi: 10.1007/s11192-014-1482-5.

[185] Michel Zitt, Alain Lelu, Martine Cadot, and Guillaume Cabanac. Science mapping and the identification of topics: Theoretical and methodological considerations. *Springer Handbook of Science and Technology Indicators*, pages 25–68, 2019. doi: 10.1007/978-3-030-02511-3_2.