Science maps for information retrieval

Bascur Cifuentes, J.P.

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

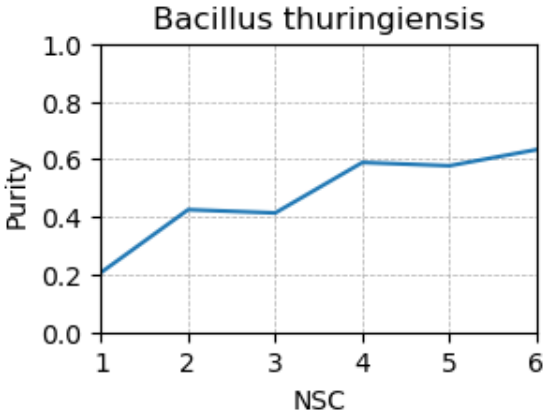

Figure 5.1: Example of a Purity profile. This is a line plot of the Purity profile of the topic *Bacillus thuringiensis [B03.510.460.410.158.218.800]* for the Policy documents BERT network calculated using Coverage 0.50. This topic has 60 topic documents among the core documents used by the Policy networks, 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

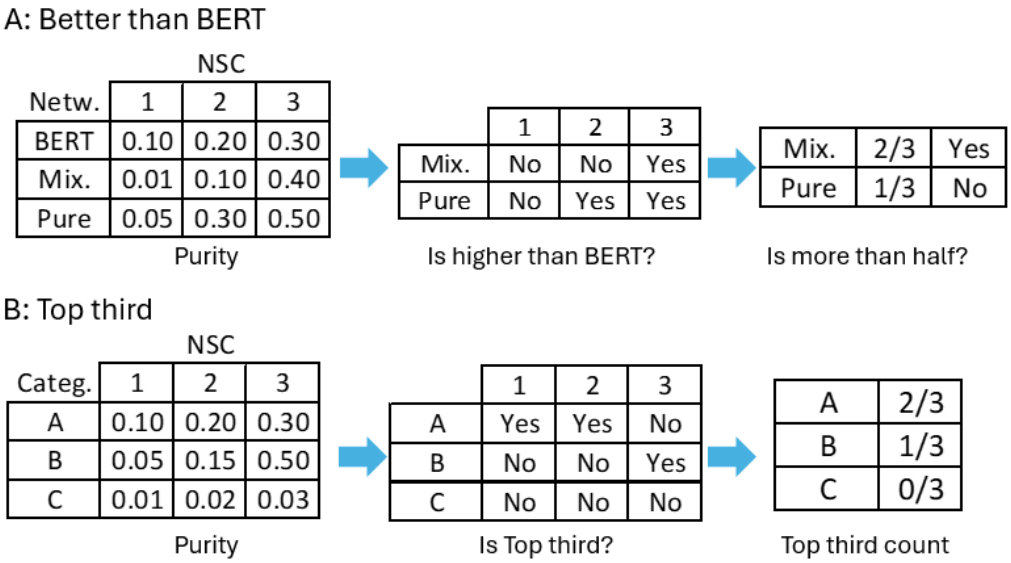

Figure 5.2: Diagram on the representation of results. A: How to calculate from topic Purity profiles if a topic has higher clustering effectiveness than BERT in the Pure or the Mixed network. In this example, a topic has higher Purity than BERT for the Mixed network, but not so for the Pure network. 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				0.2	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.				+		++					.	.		

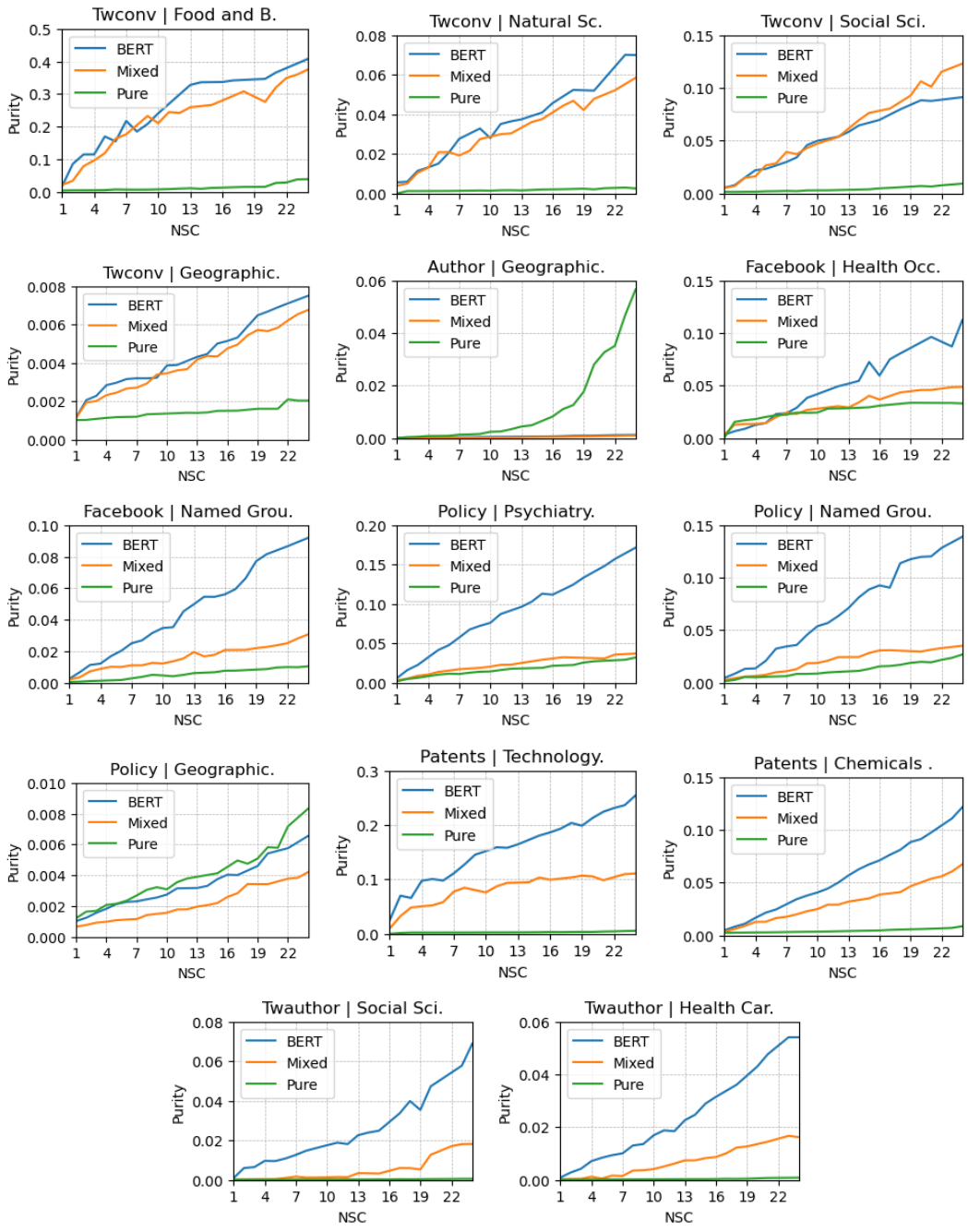

Figure 5.3: Examples of Purity of several topic categories for different networks. All profiles are for Size bin 161-320 and Coverage 0.50. To interpret these plots, it is important to keep in mind that each profile represents the average Purity and NSC across all topics in the topic category and Size 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.

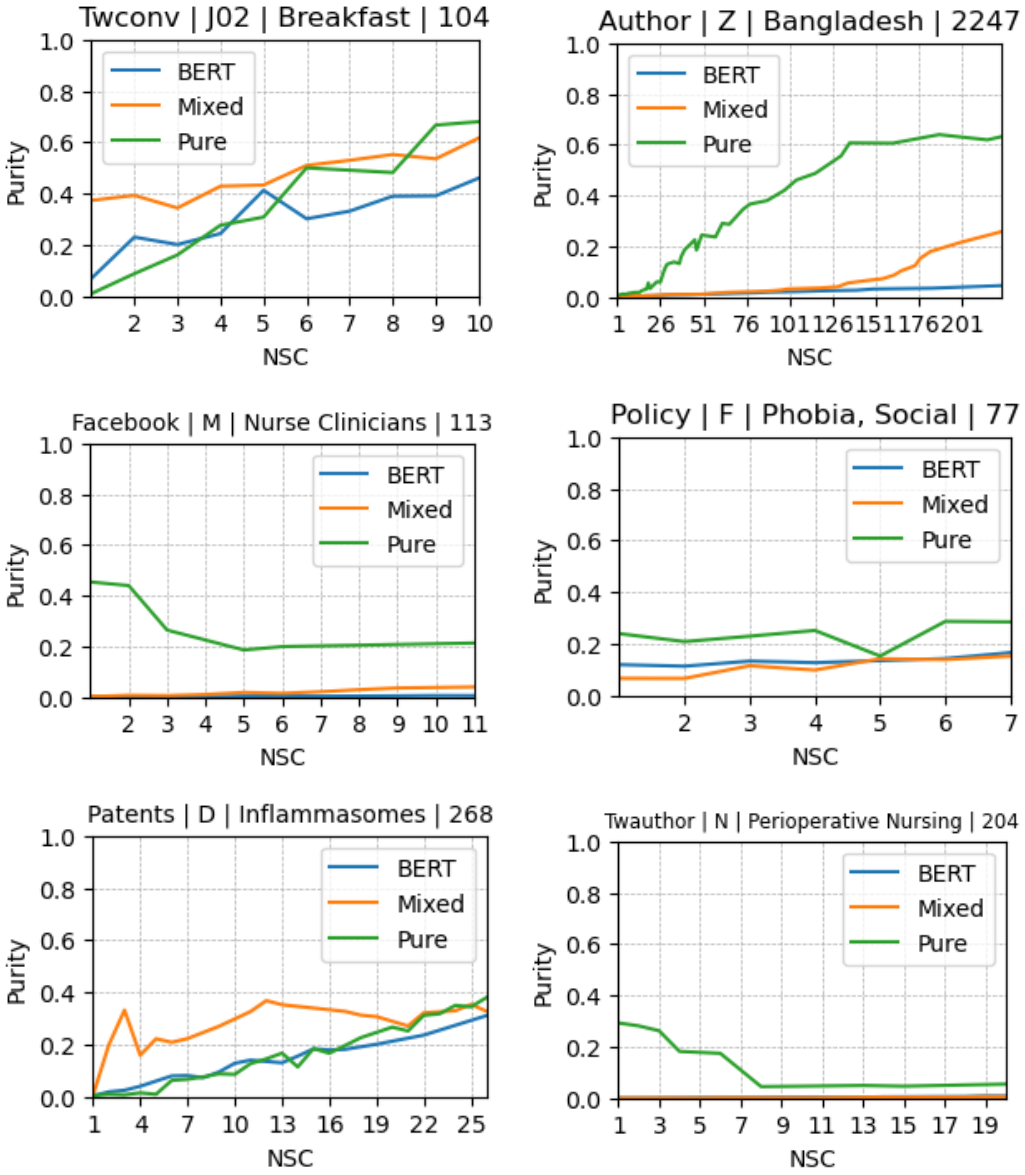

Figure 5.4: Examples of Purity profiles for individual topics across different networks. All Purity profiles are calculated for Coverage 0.50. The title of each plot indicates the external source, topic category, topic name and topic size.

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.