Bibliometric Visualization and Analysis Software: State of the Art, Workflows, and Best Practices

Michael E. Bales, Drew N. Wright, Peter R. Oxley*, and Terrie R. Wheeler*[1]

*These authors contributed equally to the work.

Abstract—Despite the demonstrated value of visualization-based modalities for measuring and mapping science, it remains common practice to search and explore the literature via databases that present lists of articles with little, if any, supplementary visual information. Identifying the desired item in a list is a familiar information retrieval paradigm with a low cognitive load. However, given the rapid emergence of the field of visual text analytics, it is time to challenge the notion that article lists should remain the dominant method to search and organize the scientific literature. One reason that visualization methods are applied relatively rarely in information retrieval may be that it is difficult to develop useful and user-friendly science mapping systems. This article summarizes key workflows for bibliometric mapping, a technique for visually representing information from scientific publications, including citation data, bibliographic metadata, and article content. It describes methods and challenges in extracting, processing, and normalizing data, reducing dimensionality, modeling topics, assigning labels, and visualizing data. It also describes software tools available to support bibliometric analysis and science mapping workflows, outlines methods from other domains that have not been widely applied in bibliometric mapping, and considers opportunities for next generation bibliometric analysis and mapping software systems.

Index Terms—bibliometrics, cognitive informatics, computer graphics, dimensionality reduction, software engineering

1 INTRODUCTION

The activity of science mapping has recently been described as "complex and unwieldy"[1], as it generally involves multiple steps which may require numerous software tools with varying levels of usability, interconnectivity, and licensing requirements. The purpose of this scoping review is to chart a way through the fog of complexity, providing readers with an overview of science mapping workflows, highlighting the strengths of the available tools, identifying pitfalls to avoid, and describing opportunities for the next generation of bibliometric analysis and science mapping software systems.

2 BACKGROUND

2.1 Science mapping

Science maps are spatial representations of how disciplines, fields, specialties, and individual documents or authors are related to one another [2]. In an effort to understand the structural and dynamic aspects of scientific research [3], [4], researchers have been creating maps of scientific domains for at least the last 50 years [5]. A variety of types of bibliographic data may be used as input, including research articles and abstracts, authors, journals, grants, and keywords or topics. Science maps may be used to convey an overview of the cognitive structure of a given field [6], [7], to determine key actors [8], to identify areas of innovation [9], to support science policy [10], [11], or to assess the evolution of scientific disciplines [12]–[16].

One type of science map, and the main focus of this article, is the bibliometric map[2], a graphical summary of a set of papers. Bibliometric maps may be derived from information about citation data, shared words or

1

- *All authors are with the Samuel J. Wood Library, Weill Cornell Medicine.*
- *Michael E. Bales e-mail: meb7002@med.cornell.edu*
- *Drew Wright e-mail: drw2004@med.cornell.edu*
- *Peter R. Oxley e-mail: pro2004@med.cornell.edu*
- *Terrie R. Wheeler e-mail: tew2004@med.cornell.edu*

[2] Bold-face words are defined in the glossary in Appendix B.

phrases, or other bibliometric elements [10]. An appealing aspect of bibliometric maps is that they share some paradigms with geographic maps; as such, they have been referred to as "landscapes of science" [11], [17]–[20]. In addition to providing an overview of the landscape for a given collection of scientific information, they also allow viewers to explore by zooming in on more information on specific sections of the landscape.

2.2 Bibliometrics

The literature on science mapping strongly overlaps with that of bibliometrics, a field concerned with measuring and analyzing science [21]–[23]. Bibliometric software systems have their own strengths and specializations; no single tool is able to support all analytical workflows [24]. Most of the more commonly used bibliometric software systems[24], [25], including the majority of those covered in this review, include features that allow users to convey data about science in a visually understandable format [5], [26].

2.3 Common workflows and analytical approaches in science mapping

While workflows used in bibliometric analysis and science mapping vary depending on the goals of the analyst, most share some common features[4], [24]. Two example science mapping analysis workflows are described in Fig. 1. Both workflows include the steps of data retrieval, pre-processing, information extraction, normalization, mapping, analysis, and visualization, after which an analyst applies domain knowledge to interpret and obtain conclusions from the results.

Three primary types of bibliometric analysis in science mapping are **citation analysis**, **co-authorship analysis**, and **co-word analysis**. Although more than one of these may be employed in a given project, they are generally used separately, as they are designed for different purposes. Methods used in citation analysis (the most common type of bibliometric analysis) [1], include **bibliographic coupling** [27], [28] and **co-citation analysis** [24], [29]. Bibliographic coupling is a method of assigning a relation between two papers which each cite the same reference. Co-citation is the frequency with which two documents are cited together.

Co-authorship analysis focuses on the relations between authors, and their affiliations, to study the social structure of collaborative networks [6], [7], [9]–[12], [30]. It may be used to assess levels of single-disciplinary and multiple disciplinary collaboration, and more generally, to examine the social and environmental factors that influence scientific collaborative behavior.

Among the three primary types of bibliometric analysis in science mapping, co-word analysis [31] is the only method that considers the semantic content of documents [1]. Co-word analysis may be applied to document titles, abstracts, or full-text documents. Co-word analysis considers the most important words or keywords in the document set, and how they co-occur, to model the conceptual structure of the text collection. The methodological foundation of co-word analysis is based on the idea that the semantic content of a set of documents is described by co-occurrences between words in the documents [32]. The term "co-word analysis" is used in the bibliometrics literature as a blanket term to cover a variety of related approaches not based strictly on co-occurrences, including **topic modeling**, which is used in workflow B shown in Fig. 1. Topic modeling can be used to assign words to topics or thematic categories. These categories can in turn be used to assign documents and labels to clusters. Some topic modeling methods can also represent semantic relations between words, based on word meaning. Topic modeling is discussed in further detail in the Results section.

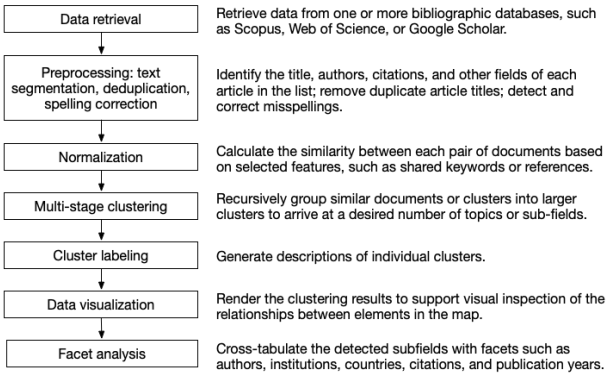
2.4 Overview of software for science mapping

General-purpose software tools that may be used for science mapping analysis and visualization include Pajek [33], R [34], UCINET [35], Gephi [36], Graphviz [37], Guess [38], Tulip [39], and Cytoscape [40] (which has become a general-purpose tool but was originally developed for use in bioinformatics research). The tools covered in this article have been developed specifically for bibliometrics analysis and science mapping. Descriptions of the more commonly-used of such tools (e.g., BibExcel, CiteSpace, IN-SPIRE, the Science of Science tool, the S&T Dynamics Toolbox, and VOSViewer) may be found in Borner et al. [25] and Cobo et al. [24], as well as in the Results section of this article.

Two types of maps are commonly used in bibliometric research, and by extension, software for bibliometric mapping: *distance-based* and *graph-based* maps [20]. In distance-based maps, the distance between two elements indicates the strength of the relation between them, with a smaller distance indicating a stronger relation [17]–[20], [41]. Graph-based maps employ a network model where elements

of information (the *nodes*, or *vertices*) are assigned *links* or *edges* when related in some way. For example, in the most common type of co-citation network produced from a collection of articles, the nodes are articles, and pairs of articles are assigned links if frequently cited together by other articles in the collection.

Workflow A



Workflow B

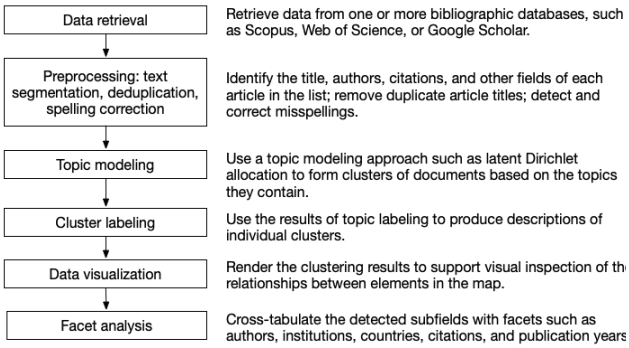


Fig. 1. Two related bibliometric clustering & science mapping analysis workflows (adapted from [21]). These are two of several common workflows used in bibliometric analysis, and are typical of approaches involving co-word analysis. In Workflow A, similarity-based calculation is used to normalize data in preparation for clustering. In Workflow B, topic modeling, rather than normalization, is used to derive clusters of documents based on themes they contain. Both of these approaches are described in further detail in the Results section.

The goal of this article is to describe common workflows and methods for bibliometric mapping and analysis, and to assess opportunities for the next generation of scientific software systems for bibliometric mapping.

3 METHODS

We devised a search strategy to identify articles that describe software systems capable of extracting and/or displaying relationships among journal articles or abstracts. We created separate queries for Scopus (http://www.scopus.com), Clarivate Analytics Web of Science (http://www.webofknowledge.com), IEEE Xplore (https://ieeexplore.ieee.org), and Engineering Village 2 (https://www.engineeringvillage.com). Specific search strategies used for each database are available in supplemental online material in Appendix A. Queries were run in December 2018. Resulting articles were combined into a single list. We performed an initial title and abstract screening to exclude articles that either did not satisfy the inclusion criteria or met one or more of the exclusion criteria (described in Table 1). After this initial review we again applied the inclusion and exclusion criteria in a review of the full text of the remaining publications, arriving at the final list of articles for inclusion. All screening was conducted using Covidence [42].

Table 1. Inclusion and exclusion criteria for article selection

Inclusion criteria	Articles describing software systems capable of extracting and/or displaying relationships among journal articles or abstracts.
Exclusion criteria	Articles describing software prototypes were excluded, as were articles covering software whose main purpose fell outside the domain of bibliometric mapping (for example, general-purpose tools for visualizing and exploring networks). Articles describing software that has not been actively supported or maintained for more than five years were also excluded, as were articles describing software mainly designed for semantic enhancement (e.g., by enriching information in articles with annotations from, or links to, other databases).

4 RESULTS

After combining the articles retrieved by the initial queries into a single list and removing duplicates, there were 728 articles. 572 of these were excluded in the initial title and abstract screening process, leaving 156 articles which were assessed for eligibility for inclusion. 126 of these were excluded because they either failed to satisfy the inclusion criteria or met one or more of the exclusion criteria (Table 1). An additional three articles were excluded because they could not be obtained either through library database subscriptions or through interlibrary loan. The remaining 27 articles were selected for inclusion. Based on a review of these articles and selected citations from among them, we selected 16 software systems for inclusion in this article (these appear in Table 2 and are summarized under "About the software systems" below).

4.1 Data retrieval

The first step in bibliometric mapping is data retrieval. Mainly due to their broad coverage and availability of downloadable publication metadata, the databases most commonly used in bibliometrics are Medline (https://www.ncbi.nlm.nih.gov/pubmed), Scopus (http://www.scopus.com), and Web of Science (http://www.webofknowledge.com). Medline, compiled by the United States National Library of Medicine, is a freely available journal citation database for data on biomedical publications. The NLM controlled vocabulary Medical Subject Headings (MeSH), are used to index citations. Scopus, Elsevier's multidisciplinary abstract and citation database, indexes journals, book series, and trade journals, and includes several measures of quality for each title. Web of Science, a citation indexing service maintained by Clarivate Analytics, is also multidisciplinary and has a broad depth of coverage. Google Scholar has been described as the world's largest academic search engine, but is used far less frequently for bibliometric analysis due to its lack of support for download of publication metadata. Additional sources of bibliographic data include arXiv (http://arxiv.org), CiteSeerX (http://citeseerx.ist.psu.edu), and ScienceDirect (http://www.sciencedirect.com). The coverage available in these databases varies by journal and by field of research, and each has its own set of advantages and disadvantages [43].

Table 2. Bibliometric mapping software systems. Some information adapted from Borner(2010) and Cobo (2011).

Software name	Developed by	Citation
BibExcel	University of Umeå (Sweden)	Persson 2009
Bibliometrix	University of Naples	Aria 2017
BiblioTools / BiblioMaps	University of Lyon	Grauwin 2011
CATAR	National Taiwan Normal University	Tseng 2013
CiteSpace	Drexel University (USA)	Chen 2006
CitNetExplorer	Leiden University	van Eck 2014
CRExplorer	Max Planck Institute	Thor 2016
Headstart	Open Knowledge Maps Team	Kraker 2019
IN-SPIRE	Pacific Northwest National Laboratory	Wise 1999
RobotReviewer	Byron Wallace, Iain Marshall, Joël Kuiper and Frank Soboczenski	Marshall 2016
S&T Dynam. Toolbox	University of Amsterdam (The Netherlands)	Leydesdorff 2019
Science of Science (Sci2) Tool	Indiana University (USA)	Sci2 Team 2009
SciMAT	University of Granada (Spain)	Cobo 2012
Utopia Documents	Lost Island Labs	Attwood 2010
VantagePoint	Search Technology, Inc.	Porter 2004
VOSViewer	Leiden University	van Eck 2009

Software name	Short description
BibExcel	System that extracts bibliographic data and outputs it in a variety of commonly-used formats (Borner 2010)
Bibliometrix	R tool for comprehensive science mapping analysis
BiblioTools / BiblioMaps	Set of scripts that create maps of science based on bibliographic data
CATAR	Software toolkit for summarizing document sets for research and strategic planning
CiteSpace	Tool that supports the use of citation patterns to analyze and visualize scientific literature
CitNetExplorer	Tool for visualizing and analyzing citation networks of scientific publications
CRExplorer	Tool to explore the historical roots of a field of research through reference publication year spectroscopy
Headstart	Software to produce knowledge maps from text, metadata, and references
IN-SPIRE	System that uses a landscape metaphor to uncover relationships, trends, and themes hidden within data
RobotReviewer	Machine learning system designed to support evidence synthesis
S&T Dynam. Toolbox	Tools for organization analysis and visualization of scholarly data (Borner 2010)
Science of Science (Sci2) Tool	System that supports the temporal, geospatial, topical, and network analysis and visualization of bibliographic collections (Aria and Cuccurullo 2017)
SciMAT	Tool that uses a longitudinal framework to support science mapping studies
Utopia Documents	PDF reader that semantically integrates visualization and data analysis tools with published research articles. (Attwood et al 2000)
VantagePoint	Text-mining tool for discovering knowledge in search results from patent and literature databases
VOSViewer	System that constructs and displays maps based on co-occurrence data

Software name	Web site	Platforms
BibExcel	https://homepage.univie.ac.at/juan.gorraiz/bibexcel/	Windows
Bibliometrix	http://www.bibliometrix.org/	All major
BiblioTools / BiblioMaps	http://www.sebastian-grauwin.com/biblio maps/	All major
CATAR	http://web.ntnu.edu.tw/~samtseng/CATAR/Readme.html	Windows
CiteSpace	http://cluster.cis.drexel.edu/~cchen/citespace/	All major
CitNetExplorer	http://www.citnetexplorer.nl/	All major
CRExplorer	http://andreas-thor.github.io/cre/	All major
Headstart	https://github.com/OpenKnowledgeMaps/Headstart	All major
IN-SPIRE	https://in-spire.pnnl.gov/	Windows
RobotReviewer	http://www.robotreviewer.net/	Online; all major
S&T Dynam. Toolbox	https://www.leydesdorff.net/software.htm	Windows
Science of Science (Sci2) Tool	https://sci2.cns.iu.edu/user/index.php	All major
SciMAT	https://sci2s.ugr.es/scimat/	All major
Utopia Documents	http://utopiadocs.com/	All major
VantagePoint	https://www.thevantagepoint.com/	Windows
VOSViewer	http://www.vosviewer.com/	All major

Software name	License type	Normalization measure
BibExcel	Freely available	Salton's cosine, Jaccard index, Vladutz and Cook measures
Bibliometrix	Freely available	Association Strength, Jaccard, Inclusion, Salton or Equivalence similarity index
BiblioTools / BiblioMaps	Freely available	Kessler's similarity (bibliographic coupling)
CATAR	Freely available for non-commercial use	Dice
CiteSpace	Freely available	Salton's cosine, dice, or Jaccard strength
CitNetExplorer	Freely available for non-commercial use	None
CRExplorer	Freely available	None
Headstart	Freely available	Cosine similarity
IN-SPIRE	Commercial software	Conditional probability
RobotReviewer	Freely available	N/A
S&T Dynam. Toolbox	Freely available and open source	Salton's cosine
Science of Science (Sci2) Tool	Freely available and open source	User defined
SciMAT	Freely available and open source	Association strength, equivalence Index, inclusion index, Jaccard's index, Salton's cosine
Utopia Documents	Freely available	N/A; links based on citations
VantagePoint	Commercial software	Pearson's r, Salton's cosine or the max proportional
VOSViewer	Freely available	Association strength

Patent and funding data may also be used in bibliometric mapping. Patent data may be retrieved from the United States Patent and Trademark Office (http://www.uspto.gov), from the European Patent Office, or from the Derwent Innovations Index provided by Clarivate Analytics Wos. Funding data may be downloaded from the National Science Foundation (http://www.nsf.gov).

Most of the systems covered in this review offer support for importing data in two or more formats, with PubMed, Scopus, and Web of Science offered most frequently. However, of the tools covered in this review, support for funding data was uncommon, with Sci2 and VantagePoint being the notable exceptions, offering support for importing NSF funding data.

4.2 Preprocessing

Data retrieved from bibliographic sources may contain errors and inconsistencies such as variations in how data are represented over time. As such, **preprocessing** is one of the most important steps in science mapping analysis. Preprocessing methods include **deduplication**, spelling correction, and time slicing [44]. **Natural language processing** (NLP) approaches may also be applied in the preprocessing step. These include **stemming** [45], **lemmatization**, and **named entity recognition** [46]. Among the tools covered in this review, SciMAT and VantagePoint are notable for the extent of their support for preprocessing. When a chosen tool does not support the type of processing or data cleanup required for an analysis, it may be necessary to carry out these steps manually or in other tools.

4.3 Normalizing data in bibliometric mapping

An essential method in distance-based bibliographic mapping is **normalization**. Normalization is an established concept in statistics but has its own distinct meaning in bibliometrics, where it refers to calculating meaningful similarities between documents [47]. This involves first defining the document features and then computing similarities between documents based on those features [21].

Similarity measures used to normalize co-occurrence data may be classified either as indirect and direct [47], or as local and global. Indirect similarity measures involve the use of co-occurrence profiles: Each object is given a vector that contains the number of co-occurrences of the object with each other object. (In co-word analysis, these objects might be word stems, recognized named entities, or multi-word phrases.) The similarity between two objects is then determined by comparing the co-occurrence profiles of the two objects. By contrast, when using direct similarity measures, the similarity between two objects is derived by calculating the number of co-occurrences of the objects, then adjusting for the total number of occurrences of each. Direct similarity measures are further classified into set-theoretic measures and

probabilistic measures [47]. Salton's cosine [48], the Jaccard index [49], the Equivalence Index [32], and the Ochiai coefficient [50] are set-theoretic measures, while **association strength** [51], [52], also known as the **proximity index** [49], [53], is a probabilistic measure. Probabilistic measures have been described as having theoretical properties that are more appropriate for normalizing co-occurrence data than set-theoretic measures [47]. Support for a variety of indirect and direct normalization approaches is widespread among the tools covered in this review. Details are in Table 2 and in the descriptions of the individual tools in the "About the Software Tools" section below.

Although they have not been cited widely in the bibliometrics literature, **neural network** based approaches may also be used to normalize data. These approaches, which operate on word vectors, include Word2vec [54], doc2vec [55], fastText [56], [57], and GloVe [58]. Word2vec and doc2vec are neural network models that represent sentences or whole documents as a vector, respectively. These methods allow for sets of articles to be clustered by similarity, based on the learned context of words and their relationships within sentences and paragraphs. Word vectors, for instance, capture the meaning of the corpus' vocabulary. Vector algebra then allows for calculation of analogous relationships between words. This can allow for surprisingly accurate computational inferences: when asked for the result of the operation "King" - "man" + "woman", one system provided the output "Queen" [59]. By extension, entire documents can also be represented and compared as vectors. The distance between documents can be found using metrics such as word mover's distance (WMD) [60], which calculates the minimum distance words needs to shift in their learned contextual space to match the words in another document. The context of the training corpus, and the type of inference being attempted, are critical factors in the quality of the outcome. Some more intriguing results are demonstrated at https://graceavery.com/word2vec-fish-music-bass/.

FastText [57] is a library designed to support scalable text classification. It employs a hierarchical classifier that organizes categories into a tree rather than a list, which improves algorithm running efficiency. FastText also employs several best practices in machine learning and natural language processing, including supplementing a bag-of-words model with subword information, and using a hidden representation to share information across classes. Another algorithm that has been shown to be effective in document clustering is GloVe [58], an unsupervised learning algorithm for obtaining vector representations for words.

Several common challenges in normalizing data must be addressed when conducting bibliometric analysis. The **polysemy** problem occurs when one term or phrase is used to represent two or more different concepts in different contexts, such as in different scientific fields (the term "normalization", as discussed above, is one such example). If this fact is not taken into account, then articles that contain these words may incorrectly be assigned a higher degree of relatedness [61]. A converse problem is the synonymy problem, where a single concept is expressed using two or more keywords; this may result in two articles receiving an artificially low similarity score [21].

Given the significant differences between various similarity measures, it is important to use the measure most appropriate for the task at hand. A detailed comparison of methods that have been applied for normalizing co-occurrence data is presented by Van Eck et al. [47].

4.4 Clustering in distance-based bibliometric analysis

In bibliometrics, article clustering refers to assigning articles into groups based on similarity. Depending upon the goals of the analyst, clustering may occur at various points in the workflow. A common approach in distance-based analysis is to use the distances calculated in the normalizing step above to group articles based on similarity[61]. Some clustering methods involve multiple stages and are thereby able to display hierarchical relationships among objects in a dataset. Hierarchical clustering [62], a common approach to multi-stage clustering, is notable in the context of science mapping, as it results in a data structure compatible with displaying topical information at varying scales. In this method, each article is at first considered a singleton cluster. The most similar pair of articles is then assigned to a cluster, in succession, until no clusters can be merged. This results in a dendrogram, which can be translated into visual cues to help with the interpretation of document groupings. Silhouette indexes can be used to determine the optimal threshold for separating groups, which can in turn improve the visual distinction between groups [63], [64].

4.6 Clustering in network-based bibliometric analysis

Data clustering algorithms for network-based approaches [51], [65]–[68], some of which are also used for distance-based approaches, include **k-means clustering** [69], Infomap [70], Louvain [71], and the Smart Local Moving Algorithm [72], [73]. **Subgroup detection** methods based on graph theory, such as a widely-used approach by Girvan and Newman [74], can also be used to assign clusters in graph-based maps. To avoid a result that includes many small clusters (e.g., clusters of just one or two articles), a minimum cluster size can be specified [61], [72].

4.7 Topic modeling

Another approach that can be applied in bibliometric analysis is topic modeling, used to discover the latent topics in a set of documents. Algorithms for topic modeling include latent semantic indexing [75], probabilistic latent semantic analysis [76], [77], and latent Dirichlet allocation (LDA) [78]. Topic analysis using LDA has been shown to improve machine learning methods for identification of relevant articles [79]. LDA assumes articles are composed of a number of "topics", and that a set of words, (e.g., a set of words within a given abstract), are representative of those topics. It can be used as an alternative to the normalization approaches described above, by forming clusters of documents based on the topics they contain. Yet despite its demonstrated utility, LDA still does not account for the contextual or semantic value of words: "cardiac", "heart" and "university" are all equally weighted and independent during analysis. This shortfall can be overcome using vector-based approaches, such as Word2vec and related methods described in the section on normalizing data, above.

Text normalization may be applied to label detected clusters of articles. Labels indicate the most important of the terms in the cluster. Text normalization assigns a weight to each term or multi-word phrase to indicate its relative importance.

4.9 Determining spatial position of visual elements

One of the steps in visualizing a bibliographic map is to determine the spatial position of visual elements. Because the calculation of similarities results in high-dimensional data that cannot readily be represented using a Cartesian coordinate system, a **dimensionality reduction** approach must be applied. Dimensionality reduction techniques such as **principal component analysis (PCA)**, **multiple correspondence analysis**, **multidimensional scaling[20], [41], [80]**, t-Distributed Stochastic Neighbor Embedding (t-SNE) [81], uniform manifold approximation and projection (UMAP), and pathfinder networks [82], [83] are widely used [1], [4], [24]. Another is the visualization of similarities mapping technique (VOS) [84], [85], which was specifically developed for use with the VOSViewer software.

Map layout for graph-based maps is typically achieved through the use of a force-directed placement technique such as the Fruchterman-Reingold [86] or Kamada-Kawai [87] algorithm. These techniques are based on physics simulations where all nodes repel one another, but linked nodes are drawn spatially proximate to one another. Another such algorithm, OpenOrd [88], is designed for networks that contain a large number of nodes (e.g., several hundred thousand). OpenOrd is open source and computationally efficient, and conveys both global and local structure.

4.10 Graphical representation of bibliometric maps

Once the spatial position of the elements of the map are determined, the map may then be rendered visually. Placing data in a visual context can help people understand its significance, revealing patterns and trends that may be more difficult to recognize in text-based data. Although significant research has focused on similarity measures [47], [89], [90] and mapping techniques [52], [91], [92], there have been fewer articles published focusing on the graphical representation of bibliometric maps [20].

Complementing map-based visualization approaches, a variety of other types of visualization approaches, such as helicocentric maps [93] and geometrical models [94] have also been applied to science maps. Additionally, some visualization approaches are designed to highlight the evolution of clusters in successive time periods, including cluster string [95]–[97], rolling clustering [67], alluvial diagrams [68], the ThemeRiver visualization [98], and thematic areas [16]. These methods are not restricted to a "map-like" visual paradigm, and can supplement science maps by conveying changes in areas of research focus over time.

4.11 Analytical methods applied to bibliometric maps

Applying analytical methods can allow for the discovery of useful information from data, networks, and maps [24]. Methods applied in science mapping include network analysis [6], [15], [99]–[101], temporal or longitudinal analysis [102], [103], geospatial analysis [104]–[106], and performance analysis [16], which aims to quantify the importance, impact, and quality of different elements of the map (e.g., the clusters) through bibliographic measures and indicators [44].

5 ABOUT THE SOFTWARE PROGRAMS

Table 2 contains a summary of all identified software tools that perform the bibliometric mapping and analysis described above. ([[Table 2 is formatted as landscape; due to limitations of Google Docs, it appears in a separate file in this folder]]). The vast majority of the systems were developed by groups at a single institution. The systems are described here alphabetically.

BibExcel [107] supports the analysis of bibliographic data, including citation and co-word analysis. It takes as input data from Web of Science, Scopus, and Procite. It supports export of data in a variety of formats, including general-purpose network analysis and visualization tools such as Pajek. Preprocessing functions include stemming, document deduplication, and elimination of low frequency items. It supports several measures for normalizing data, including Salton's cosine and Jaccard's index.

Bibliometrix [1] is a library for R that supports science mapping and bibliometrics analysis. It accepts input from Scopus and Web of Science and supports several types of citation analysis, as well as co-word analysis. Bibliometrix supports the input of data from Scopus, Web of Science, PubMed, and the Cochrane Database of Systematic Reviews. A unique feature of Bibliometrix is a plot that displays author productivity over time. It also includes support for calculating statistical measures of network structure, including average path length and various measures of centralization.

BiblioTools [108] is a set of python scripts that transforms Web of Science data to bibliometric maps, including maps based on both citation and co-word analysis. It includes functions for parsing and filtering data and requires the prior installation of several python code libraries.

CATAR [21] is a software toolkit for bibliographic clustering and mapping. It includes functions to import Web of Science data and to parse and standardize it in preparation for analysis, including deduplication. It also supports stemming, removal of stop words for co-word analysis, and document clustering and topic analysis, with document similarities calculated based on the Dice coefficient.

Citespace [109]–[111] is a Java-based system for visualizing and analyzing patterns in the scientific literature. It focuses on identifying points at which a significant innovation, such as an emerging trend or intellectual turning point, occurs in a field or domain. Although the primary import source is Web of Science data, it also reads data from PubMed, and arXiv, as well as certain grants and patent data. It supports the creation of citation-based as well as co-word networks, including automatically labeling networks using terms from articles. Supported normalization metrics include Salton's cosine, Dice, and Jaccard's index. Citespace also supports geographic maps based on the locations of authors; these are viewable in Google Earth.

CitNetExplorer [112] is a Java-based system for analyzing and visualizing citation networks. It imports data from Web of Science and allows users to identify the core literature in a given field, as well as the influence of a given author's publications on publications that are published subsequently. The system supports zooming and scrolling functionality so that users can drill down into a network to examine clusters of closely related publications. CitNetExplorer exports data to Pajek to support network-based analysis and visualization.

Cited Reference Explorer (CRExplorer) [113] is a Java-based program designed to support the identification of highly-cited papers and their influence on the historical roots of a field or researcher. It takes as input data from Web of Science or Scopus and produces time-based visualizations using reference publication year spectroscopy (RPYS).

Headstart [114] is a web-based knowledge mapping system that produces maps from text, article metadata, and references. It includes features that cluster articles and assigns labels based on keywords. Headstart is the core technology behind the Open Knowledge Maps system (https://openknowledgemaps.org).

IN-SPIRE [115] is a Windows-based visualization software system that operates from textual data, which may include journal abstracts as well as news reports, technical reports, and message traffic. The system

supports entity extraction for people, includes two main visualization approaches. The Galaxy visualization uses a metaphor of stars in the sky, with each star representing an individual document. The ThemeView visualization uses a three-dimensional terrain map to provide a high-level view of the data.

RobotReviewer [116] is a machine learning system designed to support evidence synthesis for systematic reviews. It takes as input a set of scholarly articles describing randomized controlled trials and displays information about the population, intervention, comparisons made, and the outcomes, as well as information on the study design and an assessment of risk of bias.

S&T Dynamics Toolbox [117] is a set of command line programs for bibliographic analysis, including several types of citation analysis as well as co-word analysis. It includes functions to assess collaboration at the level of institutions, cities, and countries. The toolbox does not include tools for visualization, but supports the export of data to Pajek, UCINET, and Sci2.

The Science of Science (Sci2) Tool [118] is a system that supports network-based, topical, temporal, and geospatial analysis and visualization of bibliographic data. It can import data in a wide variety of network-based and bibliographic formats, including Scopus, Web of Science, and NSF grant data. It includes functions to extract and preprocess network data and supports several types of citation-based and co-word analysis.

SciMAT [44] is a Java-based system which supports bibliographic and science mapping analysis and visualization. It allows for input of data in Web of Science and RIS format, supports both citation-based and co-word analysis, and includes several methods for normalizing data. It includes three types of visualizations: strategic diagrams, cluster networks, and evolution maps. SciMAT is noteworthy for its support of preprocessing, which features duplicate detection, identification of misspelled words, time slicing, and data reduction. Another distinguishing feature is support for the calculation of a variety of bibliometric measures based on citations, including the *h*-index.

Utopia Documents [119] is a PDF reader that enriches scholarly articles with online content. It allows readers to annotate documents, search for additional information related to article content, and view altmetrics for the article.

VantagePoint [120] is Windows-based commercial software for science mapping analysis which supports more than 190 different import formats. It supports the extraction of bibliographic metadata, including functions for data cleanup. The system supports citation-based as well as co-word analysis and visualization.

VOSviewer [20], [41] is a Java-based system for constructing and visualizing bibliometric networks based on co-citation, bibliographic coupling, or co-authorship. It also supports co-word networks. The system can import data from Web of Science, Scopus, PubMed, and RIS files. It has a visualization module with four views: label view, density view, cluster density view, and scatter view. It also supports export of data to other network-based visualization tools.

6 DISCUSSION

As the size of the base of scientific publications continues to increase, there is an ever greater need for effective methods and tools to navigate and analyze the literature. It has been said that the best we can expect from bibliometric models is a partial and imperfect reflection [121], or a "faulty mirror" of science [122]. At each step in the process of science mapping, an analyst's decisions are influenced not only by the goals of the specific analysis, but also by the idiosyncrasies of the research domain being analyzed.

That said, some steps in standard bibliometrics workflows are supported by a well-established body of literature and have coalesced into sets of best practices. For example, there is generally a shared understanding among bibliometrics practitioners of the databases used most commonly, and their main strengths and limitations. By contrast, the literature on methods to normalize bibliographic data is extensive and continues to evolve. The majority of systems employ graph-based clustering, with the exception of IN-SPIRE [115] and VOSViewer [20], [41], both of which use distance-based clustering. As mentioned earlier, probabilistic measures have theoretical properties that are more appropriate for normalizing co-occurrence data than set-theoretic measures [47]. Yet many of the tools support set-theoretic approaches to normalizing data, and under certain circumstances, expert users may have good reasons to apply them.

There is some debate in the bibliometrics community about the practicality of using full-text articles for document clustering. Using full-text articles involves at least two challenges: first, it may be difficult to obtain licences from publishers to do

research on full-text articles, and second, given their high semantic dimensionality and inconsistent formatting, it may be impractical from a methodological or computational standpoint. Given these difficulties, several researchers have recommended the compromise of using abstracts for short-text clustering [123], [124].

Short-text clustering is a recognized problem in computer science. The machine learning research community has applied a variety of methods to cluster short texts [125]–[128], including Dirichlet multinomial mixtures, global word co-occurrences, and self-aggregation [78]. Because most methods developed for clustering short texts have been developed and applied to non-scientific genres of text, such as such as news articles, they are not as effective when applied to clustering journal abstracts. As such, additional work is needed to optimize short-text clustering applied to journal abstracts.

The systems covered in this review are developed and maintained, for the most part, by individual groups at institutions, rather than large consortia. With the exception of IN-SPIRE and VantagePoint, all systems are either freely available, or freely available for non-commercial use. It is challenging to fund the development and ongoing maintenance of scientific software. Some developers have successfully sought grant funding for their systems.

The systems are diverse in terms of their main goals, interface types, and operating system support. As is common with scientific software, documentation may be sparse, and users may need to forgive difficulties in installation and unfamiliar user interfaces to accomplish tasks successfully. Given that each system has its strengths and limitations, it is not uncommon for practitioners to use different systems for different purposes. One might, for example, use one system to import and extract data, another to normalize, and another to do analysis and visualization.

To varying extents, the systems offer preprocessing tools that help clean up and standardize data. However, preprocessing steps are sometimes not applied automatically when data are imported, requiring the user to take initiative to apply them. In terms of user interfaces, although some systems produce static maps with little or no added interactive functionality, many include controls for examining specific parts of the map in additional detail [61], as by zooming and panning (moving the image around). Several of the systems have interfaces arranged in a display with several panels, such as an overview panel, a main panel, and an action panel [41]. In such a system, the entire map is displayed in a small overview panel, while the currently-viewable area of the map is displayed in a much larger main panel. Some systems that do not support visualization functionality have a multi-paneled view that excludes the visualization window. Another common user interface feature allows users to search within the articles displayed, based on title, publication year, author name, or journal name [61].

There are some well-established taxonomies of visualization approaches that can be applied in bibliometric mapping [129]–[131], but the coverage of visualization approaches, as well as the quality of implementation of the visualizations and their corresponding user interfaces, vary widely from system to system. Among the systems reviewed here, CiteSpace and VOSViewer are recognized as having visualization functions that are mature and well matched to bibliometrics and citation mapping workflows. But regardless of the bibliometrics system used, it remains common for analysts to export data to general-purpose network visualization tools such as Pajek [33], Gephi [36], or Cytoscape [40].

The next generation of bibliometrics analysis and visualization software may benefit by providing additional functions to convey changes in areas of research focus over time, such that (for example) operating a control to move forward or backward in time results in an update to the map. These changes could also be applied in the process of selecting material to include when updating a systematic review. However, this interface functionality is focused on retrospective data. To help researchers connect with important ongoing dialogues in real-time, there is an opportunity to extend bibliometrics software functionality to support emerging nontraditional data streams, such as social media and blog posts from the scientific community.

When using software systems to explore bibliometric data, analysts may face a high cognitive load when large amounts of information are presented in the display. A good strategy to limit cognitive load is to offer features that dynamically limit the amount of information that is allowed to appear in the display; expert users may be comfortable increasing the maximum number of elements that are shown in one view.

There are also opportunities to enrich bibliometrics analysis and visualization software by way of additional academic crossover with related fields. For example, a review of research on text visualization and

mining has found that existing text visualization research does not cover the majority of available text mining techniques [132]. Likewise, with the exception of CiteSpace, which displays citation counts [44], few software systems have integrated quality and impact measures (e.g., bibliographic measures, such as field-normalized citation impact), into science mapping workflows.

7 CONCLUSIONS

A wide array of software tools are available to support bibliometric analysis and science mapping workflows. Because bibliometrics is a well-established field that describes sets of methods that tend to remain stable and valid over time, it may not always be necessary to use the "latest and greatest" software system. With some level of expertise, and an awareness of the caveats of specific methods, any of the tools described in this review can be leveraged to do valid and robust bibliometrics. That said, there remain significant opportunities for the development of automated and semi-automated tools that are approachable by expert and non-expert users alike.

ACKNOWLEDGMENT

This work was supported in part by National Library of Medicine (NLM) Administrative Supplement to Clinical and Translational Science Award 5UL1TR002384-02.

REFERENCES

[1] M. Aria and C. Cuccurullo, "bibliometrix : An R-tool for comprehensive science mapping analysis," *Journal of Informetrics*, vol. 11, no. 4, pp. 959–975, Nov. 2017.

[2] H. Small, "Visualizing science by citation mapping," *J. Am. Soc. Inf. Sci.*, vol. 50, no. 9, pp. 799–813, 1999.

[3] S. A. Morris and B. Van der Veer Martens, "Mapping research specialties," *Ann. Rev. Info. Sci. Tech.*, vol. 42, no. 1, pp. 213–295, Nov. 2009.

[4] K. Börner, C. Chen, and K. W. Boyack, "Visualizing knowledge domains," *Ann. Rev. Info. Sci. Tech.*, vol. 37, no. 1, pp. 179–255, 2003.

[5] H. D. White and K. W. McCain, "Visualization of Literatures," *Annual Review of Information Science and Technology (ARIST)*, 1997.

[6] C. Weng, D. Gallagher, M. E. Bales, S. Bakken, and H. N. Ginsberg, "Understanding interdisciplinary health sciences collaborations: a campus-wide survey of obesity experts.," *AMIA Annu. Symp. Proc.*, pp. 798–802, Nov. 2008.

[7] M. E. Bales, D. R. Kaufman, and S. B. Johnson, "Evaluation of a prototype search and visualization system for exploring scientific communities.," *AMIA Annu. Symp. Proc.*, vol. 2009, pp. 24–28, Nov. 2009.

[8] R. K. Buter and E. C. M. Noyons, "Improving the functionality of interactive bibliometric science maps," *Scientometrics*, vol. 51, no. 1, pp. 55–68, Apr. 2001.

[9] P. Agarwal and D. B. Searls, "Can literature analysis identify innovation drivers in drug discovery?," *Nat. Rev. Drug Discov.*, vol. 8, no. 11, pp. 865–878, 2009.

[10] R. K. Buter, E. C. M. Noyons, M. Van Mackelenbergh, and T. Laine, "Combining concept maps and bibliometric maps: First explorations," *Scientometrics*, vol. 66, no. 2, pp. 377–387, Feb. 2006.

[11] E. C. M. Noyons, *Bibliometric Mapping as a Science Policy and Research Management Tool (PhD Thesis, Centre for Science and Technology Studies, Leiden University)*. Leiden, Netherlands: DSWO Press, 1999.

[12] M. Callon, J. Law, and A. Rip, Eds., *Mapping the dynamics of science and technology*. London: Palgrave Macmillan UK, 1986.

[13] M. Zitt and E. Bassecoulard, "Development of a method for detection and trend analysis of research fronts built by lexical or cocitation analysis," *Scientometrics*, vol. 30, no. 1, pp. 333–351, May 1994.

[14] E. C. M. Noyons, H. F. Moed, and A. F. J. Van Raan, "Integrating research performance analysis and science mapping," *Scientometrics*, vol. 46, no. 3, pp. 591–604, Nov. 1999.

[15] M. E. Bales, S. B. Johnson, J. W. Keeling, K. M. Carley, F. Kunkel, and J. A. Merrill, "Evolution of coauthorship in public health services and systems research.," *Am. J. Prev. Med.*, vol. 41, no. 1, pp. 112–117, Jul. 2011.

[16] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the Fuzzy Sets Theory field," *Journal of Informetrics*, vol. 5, no. 1, pp. 146–166, Jan. 2011.

[17] X. Polanco, C. François, and J.-C. Lamirel, "Using artificial neural networks for mapping of science and technology: A multi-self-organizing-maps approach," *Scientometrics*, vol. 51, no. 1, 2004.

[18] K. W. Boyack, B. N. Wylie, and G. S. Davidson, "Domain visualization using VxInsight® for science and technology management," *J. Am. Soc. Inf. Sci.*, vol. 53, no. 9, pp. 764–774, 2002.

[19] S. I. Fabrikant, D. R. Montello, and D. M. Mark, "The natural landscape metaphor in information visualization: The role of commonsense geomorphology," *J. Am. Soc. Inf. Sci.*, 2009.

[20] N. J. van Eck and L. Waltman, "Software survey: VOSviewer, a computer program for bibliometric mapping.," *Scientometrics*, vol. 84, no. 2, pp. 523–538, Aug. 2010.

[21] Y.-H. Tseng and M.-Y. Tsay, "Journal clustering of library and information science for subfield delineation using the bibliometric analysis toolkit: CATAR," *Scientometrics*, vol. 95, no. 2, pp. 503–528, May 2013.

[22] L. Leydesdorff, *The Challenge of Scientometrics: The Development, Measurement, and Self-organization of Scientific Communications*, Illustrated. Universal-Publishers, 2001.

[23] A. F. J. Van Raan, "Scientometrics: State-of-the-art," *Scientometrics*, vol. 38, no. 1, pp. 205–218, Jan. 1997.

[24] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma,

and F. Herrera, "Science mapping software tools: Review, analysis, and cooperative study among tools," *J. Am. Soc. Inf. Sci.*, vol. 62, no. 7, pp. 1382–1402, Jul. 2011.

[25] K. Börner, W. Huang, M. Linnemeier, R. J. Duhon, P. Phillips, N. Ma, A. M. Zoss, H. Guo, and M. A. Price, "Rete-netzwerk-red: analyzing and visualizing scholarly networks using the Network Workbench Tool," *Scientometrics*, vol. 83, no. 3, pp. 863–876, Jun. 2010.

[26] E. C. M. Noyons and A. F. J. Van Raan, "Advanced mapping of science and technology," *Scientometrics*, vol. 41, no. 1–2, pp. 61–67, Jan. 1998.

[27] S. Schiminovich, "Automatic classification and retrieval of documents by means of a bibliographic pattern discovery algorithm," *Information Storage and Retrieval*, vol. 6, no. 6, pp. 417–435, May 1971.

[28] M. M. Kessler, "Bibliographic coupling between scientific papers," *Amer. Doc.*, vol. 14, no. 1, pp. 10–25, Jan. 1963.

[29] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *J. Am. Soc. Inf. Sci.*, vol. 24, no. 4, pp. 265–269, Jul. 1973.

[30] H. P. F. Peters and A. F. J. Van Raan, "Structuring scientific activities by co-author analysis," *Scientometrics*, vol. 20, no. 1, pp. 235–255, Jan. 1991.

[31] M. Callon, J. P. Courtial, W. A. Turner, and S. Bauin, "From translations to problematic networks: An introduction to co-word analysis," *Social Science Information*, vol. 22, no. 2, pp. 191–235, Mar. 1983.

[32] M. Callon, J. P. Courtial, and F. Laville, "Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemsitry," *Scientometrics*, vol. 22, no. 1, pp. 155–205, Sep. 1991.

[33] V. Batagelj and A. Mrvar, "Pajek-program for large Network analysis," Jan. 1998.

[34] R. Ihaka and R. Gentleman, "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, vol. 5, no. 3, pp. 299–314, Sep. 1996.

[35] S. P. Borgatti, M. G. Everett, and L. C. Freeman, "UCINET 6 for Windows: Version 6.199."

[36] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks," Mar. 2009.

[37] Graphviz Team, *Graphviz – Graph Visualization Software*. AT&T Research Group, 2019.

[38] E. Adar, *Guess: The Graph Exploration System*. cond.org, 2019.

[39] D. Auber, "Tulip — A huge graph visualization framework," in *Graph Drawing Software*, M. Jünger and P. Mutzel, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 105–126.

[40] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks.," *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003.

[41] N. J. van Eck and L. Waltman, "Vosviewer: A Computer Program for Bibliometric Mapping," *ERIM*, 2009.

[42] Veritas Health Innovation, *Covidence systematic review software*. Melbourne, Australia: Veritas Health Innovation, 2019.

[43] M. E. Falagas, E. I. Pitsouni, G. A. Malietzis, and G. Pappas, "Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses," *FASEB J.*, vol. 22, no. 2, pp. 338–342, Feb. 2008.

[44] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "SciMAT: A new science mapping analysis software tool," *J. Am. Soc. Inf. Sci.*, vol. 63, no. 8, pp. 1609–1630, Aug. 2012.

[45] M. Anjali and G. Jivani, "A comparative study of stemming algorithms," *Int. J. Comp. Tech. Appl*, vol. 2, no. 6, pp. 1930–1938, 2011.

[46] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *LI.*, vol. 30, no. 1, pp. 3–26, Aug. 2007.

[47] N. J. van Eck and L. Waltman, "How to normalize cooccurrence data? An analysis of some well-known similarity measures," *J. Am. Soc. Inf. Sci.*, vol. 60, no. 8, pp. 1635–1651, Aug. 2009.

[48] G. Salton and M. J. McGill, *Introduction to modern information retrieval*. McGraw-Hill, 1983.

[49] H. P. F. Peters and A. F. J. van Raan, "Co-word-based science maps of chemical engineering. Part I: Representations by direct multidimensional scaling," *Research Policy*, vol. 22, no. 1, pp. 23–45, Feb. 1993.

[50] Q. Zhou and L. Leydesdorff, "The normalization of occurrence and co-occurrence matrices in bibliometrics using cosine similarities and Ochiai coefficients," *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 11, pp. 2805–2814, Nov. 2016.

[51] N. Coulter, I. Monarch, and S. Konda, "Software engineering as seen through its research literature: A study in co-word analysis," *J. Am. Soc. Inf. Sci.*, vol. 49, no. 13, pp. 1206–1223, 1998.

[52] N. J. Van Eck and L. Waltman, "Bibliometric mapping of the computational intelligence field," *Int. J. Unc. Fuzz. Knowl. Based Syst.*, vol. 15, no. 05, pp. 625–645, Oct. 2007.

[53] A. Rip and J. P. Courtial, "Co-word maps of biotechnology: An example of cognitive scientometrics," *Scientometrics*, vol. 6, no. 6, pp. 381–400, Nov. 1984.

[54] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space," *Google Research*, Jan. 2013.

[55] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," *PMLR*, vol. 32, pp. 1188–1196, Jun. 2014.

[56] "Release v0.1.0 · facebookresearch/fastText · GitHub." [Online]. Available: https://github.com/facebookresearch/fastText/. [Accessed: 14-Aug-2019].

[57] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of Tricks for Efficient Text Classification," presented at the Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2016.

[58] J. Pennington, R. Socher, and C. Manning, "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*,

Stroudsburg, PA, USA, 2014, pp. 1532–1543.
[59] T. Mikolov, W. Yih, and G. Zweig, "Linguistic Regularities in Continuous Space Word Representations.," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, vol. 13, pp. 746–751.
[60] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From Word Embeddings To Document Distances," *PMLR*, vol. 37, pp. 957–966, Jul. 2015.
[61] N. J. van Eck and L. Waltman, "Citation-based clustering of publications using CitNetExplorer and VOSviewer.," *Scientometrics*, vol. 111, no. 2, pp. 1053–1070, Feb. 2017.
[62] Salton and Gerald, *Automatic text processing*. Addison-Wesley Longman Publishing Co., Inc., 1988.
[63] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987.
[64] P. Ahlgren and B. Jarneving, "Bibliographic coupling, common abstract stems and clustering: A comparison of two document-document similarity approaches in the context of science mapping," *Scientometrics*, vol. 76, no. 2, pp. 273–290, Aug. 2008.
[65] C. Chen, F. Ibekwe-SanJuan, and J. Hou, "The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis," *J. Am. Soc. Inf. Sci.*, vol. 61, no. 7, pp. 1386–1409, Mar. 2010.
[66] P. Chen and S. Redner, "Community structure of the physical review citation network," *Journal of Informetrics*, vol. 4, no. 3, pp. 278–290, Jul. 2010.
[67] V. Kandylas, S. P. Upham, and L. H. Ungar, "Analyzing knowledge communities using foreground and background clusters," *ACM Trans. Knowl. Discov. Data*, vol. 4, no. 2, pp. 1–35, May 2010.
[68] M. Rosvall and C. T. Bergstrom, "Mapping change in large networks.," *PLoS ONE*, vol. 5, no. 1, p. e8694, Jan. 2010.
[69] D. J. Mackay, *Information Theory, Inference and Learning Algorithms*. Cambridge, UK: Cambridge University Press, 2003.
[70] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure.," *Proc Natl Acad Sci USA*, vol. 105, no. 4, pp. 1118–1123, Jan. 2008.
[71] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *J. Stat. Mech.*, vol. 2008, no. 10, p. P10008, Oct. 2008.
[72] L. Waltman and N. J. van Eck, "A new methodology for constructing a publication-level classification system of science," *J. Am. Soc. Inf. Sci.*, vol. 63, no. 12, pp. 2378–2392, Dec. 2012.
[73] L. Waltman and N. J. van Eck, "A smart local moving algorithm for large-scale modularity-based community detection," *Eur. Phys. J. B*, vol. 86, no. 11, p. 471, Nov. 2013.
[74] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks.," *Proc Natl Acad Sci USA*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.
[75] C. H. Papadimitriou, P. Raghavan, H. Tamaki, and S. Vempala, "Latent semantic indexing: A probabilistic analysis," *Journal of Computer and System Sciences*, vol. 61, no. 2, pp. 217–235, Oct. 2000.
[76] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Am. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, Sep. 1990.
[77] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge.," *Psychol. Rev.*, vol. 104, no. 2, pp. 211–240, 1997.
[78] Q. Jipeng, Q. Zhenyu, L. Yun, Y. Yunhao, and W. Xindong, "Short Text Topic Modeling Techniques, Applications, and Performance: A Survey," *arXiv*, Apr. 2019.
[79] Y. Mo, G. Kontonatsios, and S. Ananiadou, "Supporting systematic reviews using LDA-based document representations.," *Syst. Rev.*, vol. 4, p. 172, Nov. 2015.
[80] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, Mar. 1964.
[81] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, Nov. 2008.
[82] R. W. Schvaneveldt, F. T. Durso, and D. W. Dearholt, "Network structures in proximity data," vol. 24, Elsevier, 1989, pp. 249–284.
[83] A. Quirin, O. Cordón, J. Santamaría, B. Vargas-Quesada, and F. Moya-Anegón, "A new variant of the Pathfinder algorithm to generate large visual science maps in cubic time," *Information Processing & Management*, vol. 44, no. 4, pp. 1611–1623, Jul. 2008.
[84] L. Waltman, N. J. van Eck, and E. C. M. Noyons, "A unified approach to mapping and clustering of bibliometric networks," *Journal of Informetrics*, vol. 4, no. 4, pp. 629–635, Oct. 2010.
[85] N. J. van Eck and L. Waltman, "VOS: A new method for visualizing similarities between objects," in *Advances in data analysis*, R. Decker and H.-J. Lenz, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 299–306.
[86] T. M. J. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Softw: Pract. Exper.*, vol. 21, no. 11, pp. 1129–1164, Nov. 1991.
[87] T. Kamada and S. Kawai, "An algorithm for drawing general undirected graphs," *Inf. Process. Lett.*, vol. 31, no. 1, pp. 7–15, Apr. 1989.
[88] S. Martin, W. M. Brown, R. Klavans, and K. W. Boyack, "OpenOrd: an open-source toolbox for large graph layout," in *Visualization and Data Analysis 2011*, 2011, vol. 7868, p. 786806.
[89] P. Ahlgren, B. Jarneving, and R. Rousseau, "Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient," *J. Am. Soc. Inf. Sci.*, vol. 54, no. 6, pp. 550–560, Apr. 2003.
[90] R. Klavans and K. W. Boyack, "Identifying a better measure of relatedness for mapping science," *J. Am. Soc. Inf. Sci.*, vol. 57, no. 2, pp. 251–263, Jan. 2006.
[91] K. W. Boyack, R. Klavans, and K. Börner, "Mapping the backbone of science," *Scientometrics*, vol. 64, no.

3, pp. 351–374, Aug. 2005.
[92] H. D. White, "Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists," *J. Am. Soc. Inf. Sci.*, vol. 54, no. 5, pp. 423–434, Mar. 2003.
[93] F. de Moya-Anegón, B. Vargas-Quesada, Z. Chinchilla-Rodriguez, E. Corera-Álvarez, V. Herrero-Solana, and F. J. Munoz-Fernández, "Domain analysis and information retrieval through the construction of heliocentric maps based on ISI-JCR category cocitation," *Information Processing & Management*, vol. 41, no. 6, pp. 1520–1533, Dec. 2005.
[94] A. Skupin, "Discrete and continuous conceptualizations of science: Implications for knowledge domain visualization," *Journal of Informetrics*, vol. 3, no. 3, pp. 233–245, Jul. 2009.
[95] H. Small, "Tracking and predicting growth areas in science," *Scientometrics*, vol. 68, no. 3, pp. 595–610, Sep. 2006.
[96] H. Small and P. Upham, "Citation structure of an emerging research area on the verge of application," *Scientometrics*, vol. 79, no. 2, pp. 365–375, May 2009.
[97] S. P. Upham and H. Small, "Emerging research fronts in science and technology: patterns of new knowledge development," *Scientometrics*, vol. 83, no. 1, pp. 15–38, Apr. 2010.
[98] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "ThemeRiver: visualizing thematic changes in large document collections," *IEEE Trans. Vis. Comput. Graph.*, vol. 8, no. 1, pp. 9–20, 2002.
[99] P. J. Carrington, J. Scott, and S. Wasserman, Eds., *Models and Methods in Social Network Analysis*. Cambridge University Press, 2005.
[100] D. J. Cook and L. B. Holder, Eds., *Mining Graph Data*, Illustrated. John Wiley & Sons, 2006.
[101] S. Wasserman and K. Faust, *Social Network Analysis*. Cambridge: Cambridge University Press, 1994.
[102] E. Garfield, "Scientography: Mapping the tracks of science," *Current Contents: Social & Behavioral Sciences*, vol. 7, no. 45, pp. 5–10, 1994.
[103] D. de Solla Price and S. Gürsey, "Studies in Scientometrics I Transience and Continuance in Scientific Authorship," *Ciência da Informação*, 1975.
[104] L. Leydesdorff and O. Persson, "Mapping the geography of science: Distribution patterns and networks of relations among cities and institutes," *J. Am. Soc. Inf. Sci.*, 2010.
[105] M. Batty, "The geography of scientific citation," *Environ. Plan. A*, vol. 35, no. 5, pp. 761–765, May 2003.
[106] H. Small and E. Garfield, "The geography of science: disciplinary and national mappings," *Journal of Information Science*, vol. 11, no. 4, pp. 147–159, Oct. 1985.
[107] O. Persson, R. Danell, and J. W. Schneider, "Celebrating Scholarly Communication Studies: A Festschrift for Olle Persson at his 60th Birthday ," *Special volume of the e-newsletter of the international society for scientometrics and informetrics*, pp. 9–24, 2009.
[108] S. Grauwin and P. Jensen, "Mapping scientific institutions," *Scientometrics*, vol. 89, no. 3, pp. 943–954, Dec. 2011.
[109] K. Allendoerfer, S. Aluker, G. Panjwani, J. Proctor, D. Sturtz, M. Vukovic, and Chaomei Chen, "Adapting the cognitive walkthrough method to assess the usability of a knowledge domain visualization," in *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, 2005, pp. 195–202.
[110] C. Chen, *CiteSpace: A Practical Guide for Mapping Scientific Literature*, Illustrated, Reprint. Nova Science Publishers.
[111] C. Chen, "CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature," *J. Am. Soc. Inf. Sci.*, vol. 57, no. 3, pp. 359–377, Feb. 2006.
[112] N. J. van Eck and L. Waltman, "CitNetExplorer: A new software tool for analyzing and visualizing citation networks," *Journal of Informetrics*, vol. 8, no. 4, pp. 802–823, Oct. 2014.
[113] A. Thor, W. Marx, L. Leydesdorff, and L. Bornmann, "Introducing CitedReferencesExplorer (CREXplorer): A program for reference publication year spectroscopy with cited references standardization," *Journal of Informetrics*, vol. 10, no. 2, pp. 503–515, May 2016.
[114] P. Kraker, C. Kittel, M. Schramm, R. Bachleitner, T. Arrow, S. Chamberlain, A. Enkhbayar, Y. Stein, P. Weissensteiner, M. Skaug, K. Leinweber, and Open Knowledge Maps team and contributors, *Headstart*. Zenodo, 2019.
[115] J. A. Wise, "The ecological approach to text visualization," *J. Am. Soc. Info. Sci.*, vol. 50, no. 13, pp. 1224–1233, 1999.
[116] I. J. Marshall, J. Kuiper, and B. C. Wallace, "RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials.," *J. Am. Med. Inform. Assoc.*, vol. 23, no. 1, pp. 193–201, Jan. 2016.
[117] L. Leydesdorff, *Software and data of Loet Leydesdorff*. University of Amsterdam, 2019.
[118] Sci2 Team, *Science of Science (Sci2) Tool*. Indiana University and SciTech Strategies, 2009.
[119] T. K. Attwood, D. B. Kell, P. McDermott, J. Marsh, S. R. Pettifer, and D. Thorne, "Utopia documents: linking scholarly literature with research data," *Bioinformatics*, vol. 26, no. 18, pp. i568-74, Sep. 2010.
[120] "Home - The VantagePoint." [Online]. Available: https://www.thevantagepoint.com/. [Accessed: 22-Mar-2019].
[121] H. Small, "Paradigms, citations, and maps of science: A personal history," *J. Am. Soc. Inf. Sci.*, vol. 54, no. 5, pp. 394–399, Mar. 2003.
[122] B. C. Griffith, "Science literature—how faulty a mirror of science?," *AP*, vol. 31, no. 8, pp. 381–391, Aug. 1979.
[123] C. Li, J. Guo, Y. Lu, J. Wu, Y. Zhang, Z. Xia, T. Wang, D. Yu, X. Chen, and P. Liu, "LDA meets word2vec: A novel model for academic abstract clustering," in *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*, New York, New York, USA, 2018, pp. 1699–1706.
[124] P. Makagonov, M. Alexandrov, and A. Gelbukh, "Clustering abstracts instead of full texts," in *Text, speech and dialogue*, vol. 3206, P. Sojka, I. Kopeček, and K. Pala, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 129–135.
[125] C. T. Zheng, C. Liu, and H. S. Wong, "Corpus-based

topic diffusion for short text clustering," *Neurocomputing*, vol. 275, pp. 2444–2458, Jan. 2018.
[126] L. Cagnina, M. Errecalde, D. Ingaramo, and P. Rosso, "An efficient Particle Swarm Optimization approach to cluster short texts," *Inf Sci (Ny)*, vol. 265, pp. 36–49, May 2014.
[127] S. Seifzadeh, A. K. Farahat, M. S. Kamel, and F. Karray, "Short-Text Clustering using Statistical Semantics," in *Proceedings of the 24th International Conference on World Wide Web - WWW '15 Companion*, New York, New York, USA, 2015, pp. 805–810.
[128] J. Xu, P. Wang, G. Tian, B. Xu, J. Zhao, F. Wang, and H. Hao, "Short Text Clustering via Convolutional Neural Networks," presented at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies, 2015.
[129] A. B. Alencar, M. C. F. de Oliveira, and F. V. Paulovich, "Seeing beyond reading: a survey on visual text analytics," *WIREs Data Mining Knowl Discov*, vol. 2, no. 6, pp. 476–492, Nov. 2012.
[130] K. Kucher and A. Kerren, "Text visualization techniques: Taxonomy, visual survey, and community insights," in *2015 IEEE Pacific Visualization Symposium (PacificVis)*, 2015, pp. 117–121.
[131] P. Federico, F. Heimerl, S. Koch, and S. Miksch, "A survey on visual approaches for analyzing scientific literature and patents.," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 9, pp. 2179–2198, 2017.
[132] S. Liu, X. Wang, C. Collins, W. Dou, F. Ouyang, M. El-Assady, L. Jiang, and D. A. Keim, "Bridging Text Visualization and Mining: A Task-Driven Survey.," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 7, pp. 2482–2504, Jul. 2019.

Michael E. Bales is the Research Impact and Evaluation Informationist at the Samuel J. Wood Library at Weill Cornell Medicine, where his work focuses on data curation, research impact assessment, and informatics policy. He is a grant writer and lecturer in Team Science for Weill Cornell's Clinical and Translational Science Center. Michael obtained his PhD in Biomedical Informatics at Columbia University, his MPH in Epidemiology at the University of Minnesota, and completed a Public Health Informatics Fellowship at the U.S. Centers for Disease Control and Prevention.

Drew N. Wright is the Scholarly Communications Librarian at Weill Cornell Medical Library, where he serves as a liaison between the library and the research community and provides support to students and faculty regarding publishing, grant-writing, experimental design, and data management. Prior to that he served as a research assistant at New York University and Northeastern University. Drew has a B.S. in Chemical Engineering (2003) and a M.S. in Library and Information Science (2012). His current research interests include systematic reviews, bibliometrics, and scientific data curation.

Peter R. Oxley is the Associate Director of Research Services at the Samuel J. Wood Library & C.V. Starr Biomedical Information Center. In this capacity, he oversees provisioning and education in research applications, databases and services. This includes managing the institutional Data Core, a secure enclave cloud computing environment for storage and analysis of clinical data; establishing the library bioinformatics service, teaching and consulting in data science and bioinformatics; and managing the Scientific Software Hub for discovery and provisioning of software licensing across the institution. His research interests include data literacy, scientific reproducibility, machine learning, and bioinformatics. Peter has a PhD in Behavioral Genetics and Evolution from the University of Sydney, a graduate diploma in Genetic Counseling from Newcastle University, and a bachelor's in Molecular Biology and Genetics (with honors) from the University of Sydney.

Terrie R. Wheeler, AMLS, has over 37 years of experience in librarianship; primarily in leadership positions. Her passion for developing innovative, high-performance teams has revolutionized the future of three libraries during her career. Her keen interest in information technology and in communicating the value of libraries in language with data that administrators understand has led her to champion development of automated open source tools to support the work of information professionals. As the current Director of the Samuel J. Wood Library and C.V. Starr Biomedical Information Center, she has made exploring and funding the development of open source tools part of her portfolio. Wheeler holds a Master of Library Science from the University of Michigan.