Mapping Scientific Research

Group 2

Alexander Grass, Lea Novak, and Danica Radulovic

Institute of Interactive Systems and Data Science (ISDS),
Graz University of Technology
A-8010 Graz, Austria

24 May 2019

Abstract

Science mapping is essential to an understanding of the structure and dynamics of scientific knowledge. This survey provides an introduction into the topic - what it is, how it is done and what tools to use in practice. An overview of the common methods like citation, co-citation, co-authorship, bibliographic coupling and co-occurrence of words is given. In addition, evaluations and discussions about the most common tools like SciMat, Sci2, VOSviewer and CitNetExplorer are provided. The focus of this survey is not on the content of a particular field, but on the implications, necessities and possibilities science mapping brings in general.

Contents

List of Figures

Chapter 1

Introduction

Science Mapping is a way of visualizing the structure of science. It provides an insight into the underlying dynamics and the interconnection between the different areas and fields. Historically, science mapping emerged from traditional library information science and computer science [Chen et al. 2017].

1.1 Motivation

There are different motivators for science mapping which lead to a big interest into the field. The most important are the ease of research and drawing a "landscape of science".

1.1.1 Research

Science is growing ever faster. It is hard for researchers to cope with the speed science is moving on. It is often the desire of a researcher to explore a specific area of science in depth and read as much literature as possible about related topics. There are various approaches to go along, like for example information retrieval systems. These systems are the conventional approach to solve the researcher's problem. Science mapping on the other hand serves as an alternative approach. A co-citation network for example could also reveal much information on what is related to what and even how important a specific publication is to the scientific community.

1.1.2 The Big Picture

It is very important to visualize the evolution of science and answer questions like:

- How are the big fields of science related to each other?

- Where did the fields emerge from?

- How massive is a field?

Science mapping provides an answer to those questions in a visualized form and in a way which is easy to grasp for the spectator. Most of the approaches are focused on identifying clusters of authors, papers or references, but there are also alternative methods based on co-word analysis to identify semantic schemes [K. W. Boyack et al. 2005]. Many science mapping systems offer the possibility to interact with the visualization, which brings further understanding of the data.

1.2 The Measurement

The most appropriate unit of measurement in the scientific field are citations. A citation is always an indicator of relevance, regardless of the cited source being supportive or dismissive [Chen et al. 2017]. If there is a citation, then the source cannot be unimportant. This neutral character is what makes

citation the most popular unit of measurement. However, there are some scenarios in which other units of measurements might be better or more appropriate, like word occurrence or co-occurrence, bibliographic coupling or co-authorship.

Chapter 2

Methods

There are three main methods used for science mapping. Although, this survey is mostly based on co-citation analysis, one can find also brief explanations of co-authorship analysis and bibliographic coupling.

2.1 Co-Authorship Analysis

In case when two scientists appear as authors of the same publication, the co-authorship link is built between them. The more times those two scientists share the authorship of a paper, the stronger is the co-authorship connection between them. Therefore, co-authorship analysis measures the frequency of co-occurrence of two authors. As an example in Figure 2.1 can be seen two authors A and B that would have co-authorship link of weight two between them, since they co-occur as authors of 2 papers. Furthermore, co-authorship can be measured between two groups of authors or two institutions.

Currently this method is often used in science as a proxy for research collaboration. However, according to Ponomariov [Ponomariov and Boardman 2016], this should not always be considered as the most reliable method of science mapping, since it happens quite often that a person is listed as an author on a particular publication for some reason other than research collaboration.

2.2 Bibliographic Coupling

Two publications are bibliographically coupled when they both have a common third work on their reference list. The strength of this coupling increases with a larger number of common references. As an example in Figure 2.2 When papers C, D and E appear in both reference lists of papers A and B, the bibliographic coupling link between papers A and B would have a weight of three. The idea of bibliographic coupling is to measure the similarity of subjects of a pair of publications. This method was firstly introduced by M.M.Kessler in 1963 [Kessler 1963] and mostly focuses on groups of papers that cite one source publication.

2.3 Co-Citation Analysis

The idea of co-citation was brought up by Irina Marshakova and Henry Small in 1973. This method is considered as the most effective and efficient way of science mapping. If two papers appear together in a reference list of the third paper, then these two papers have a co-citation relationship. The more co-citations two publications receive, the higher is the weight of a link between them, frequently called co-citation strength. Co-citation coupling is extended in the following years to journals and authors [Colavizza et al. 2018]. Therefore, there are three different levels of classification of co-citation, depending on what is the object of analysis [Egghe and Rousseau 2002]:
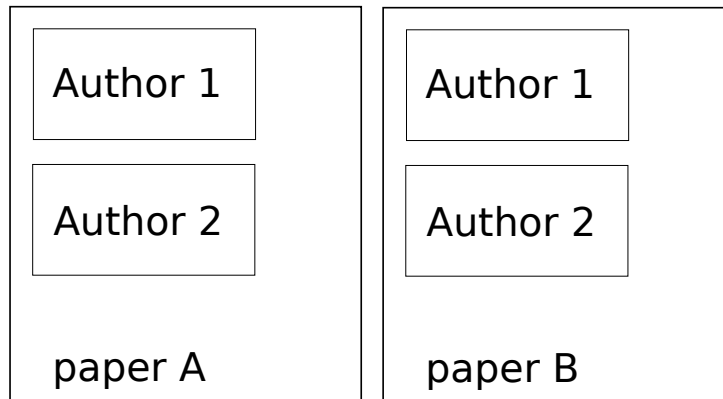
Figure 2.1: Co-authorship of A and B: Two Scientists A and B both occur as co-authors of both paper 1 and paper 2. [Drawn by the authors of this survey.]
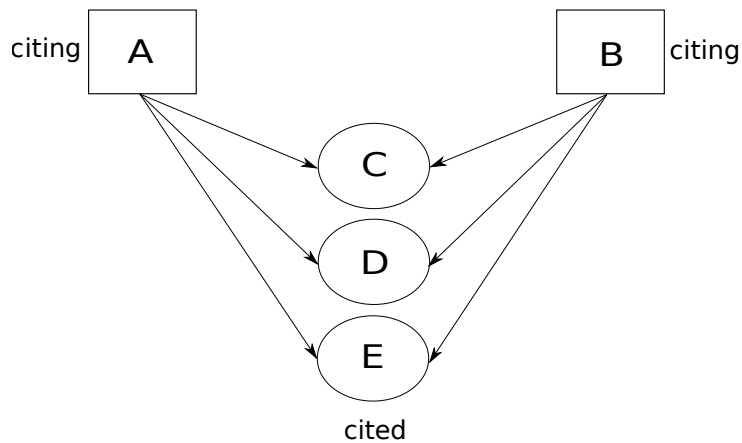
Figure 2.2: Bibliographic coupling of A and B: Both Papers A and B have papers C,D and E in their reference lists. [Redrawn by the authors of this survey. Original by: [Egghe and Rousseau 2002]]

- Documents co-citation is used when search on similar documents is conducted. This was the original idea of co-citing.

- Author co-citation, extended by White and Griffit in 1981 [White and Griffith 1981] has been used for conducting analysis of intellectual structure of science.

- Journal co-citation, extended by McCain in 1991 [McCain 1991], is used for analysing journals and creating collections that are oriented to the similar research topics.

Figure 2.3: Co-citation link between papers A and B of weight 3 exists since they are both on the reference lists of papers C,D and E. [Redrawn by the authors of this survey. Original by: [Egghe and Rousseau 2002]]

Chapter 3
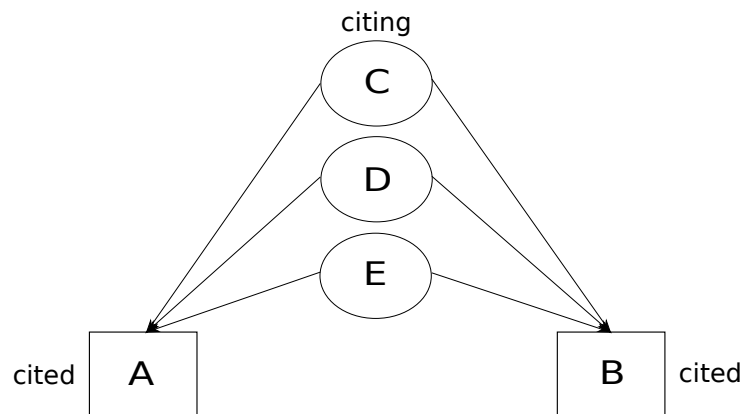
Science Citation Index and Data Sources

3.1 Science Citation Index

For getting an access to the knowledge of science it was necessary to study science as a system first, recognize the intellectual structure of science, and be able to move through this structured data through the time. The idea of creating an index for the whole science, which is based on citations and therefore creates connections using references was carried out by Eugene Garfield, specialist in the world of library and information science [Small 2018]. In this way the Science Citation Index(SCI) was launched in 1964. Originally SCI was created as an index for citations between papers published in academic journals since 1900. This commercial database is currently owned by Clarivate Analytics [Clarivate Analytics 2019a]. SCI is divided into two parts:

- Source Author Index, which includes the information what each scientist has published.

- Citation Index, which has purpose to identify where and how often papers of the certain author are cited.

SCI is meant to help scientists to accelerate their researches, by speeding up the search for the necessary data and evaluate the impact of their own work. The web version of SCI is called SCI Expanded and according to the current update in the beginning of May 2019 [Nguyen Tat Thang 2019] it covers 9148 journals.

3.2 Data Sources

There are two very large data sources which provide metadata relying on co-citation coupling and provide the simple search after publications and its authors. These two databases are huge competitors, however, none of these is a complete dataset. Therefore they complement each other well.

3.2.1 Web of Science

Web of Science(WOS) [Clarivate Analytics 2019b] is a large data source which covers over 73 million records, owned by Clarivate Analytics [Clarivate Analytics 2019a]. Its core collection [Clarivate Analytics 2019c] contains information from more that 18000 journals and consists of:

- Science Citation Index Expanded, which covers more than 9000 journals across 150 different disciplines from 1900 to present,

- Social Sciences Citation Index which indexes 3200 journals across 55 social science disciplines from 1900 to present,

- Arts and Humanities Citation Index that covers more than 1700 journals from 1975 to present,

- Emerging Sources Citation Index which covers over 5000 journals including scientific, social science and humanities beyond the high-impact literature, mentioned in the previous ones,

- Book Citation Index indexes over 80000 books from 2005 to present, but increasing by adding 10000 new books each year,

- Conference Proceedings Citation Index includes more than 180000 conference proceedings from 1990 to present.

For each paper in the WOS core collection the information about all authors, all author affiliations, the abstract and keywords, all the cited references and funding acknowledgments is captured. Additionally the metadata can be added (e.g. ORCID identifiers).

Accessing WOS is possible through the library or an institutional login. Each user can register himself and use his account for saving searches, lists of records and can set up search and citation alerts.

WOS provides three searching options: basic, cited reference and advanced search. In advanced search an extended version of search operators is provided: AND, OR, NOT, NEAR(to find records containing wanted terms within a certain number of words of each other). Search results can then be sorted by: times cited, recently added, publication date, source, first author and conference name. Furthermore, results can be refined by setting various filters and citation reports can be created for any set of results with less than 10000 records, which visually presents founded results.

Figure 3.1 shows a Web Of Science dataset about Information Visualization. Usually, the dataset is downloaded as tab-delimited plain text file, but can also be exported to endnote, bibtex, HTML, and sent to InCites, Refworks, Publons or sent via email. The first row contains various tags like PT (publication type), AU (authors), BA (book authors), TI (document title), AB (abstract), PU (publisher), CT (conference title), and many more. A more detailed list can be found at [Clarivate 2018]. The next row contains the first document entry, where each column is separated with tabs. This dataset contains a "J" at the beginning and stands for journal. There exist various others like B (book), J (journal), P (patent), and S (book in series). Next are authors displayed by surname and firstname and separated again by a tab followed by the next metadata entry. A new document always begins with a newline.

3.2.2 Scopus

Scopus [Elsevier 2019b] is another large database for citations of scientific journals, books and conference proceedings, containing over 69 million records. Scopus is owned by Elsevier [Elsevier 2019a] and launched in 2004. Scopus uses different metrics for different levels (author, journal or publication) in order to provide the best search results. Some of the metrics are:

- h-index - an author-level metric which manages to measure productivity and citation impact of the publication of a scientist),

- CiteScore - released by Elsevier in 2016, it is a very accurate metric, which is calculated every year using the metrics from the last 3 years, in a way that ranks journals using the Scopus citation data,

- SJR (SCImago Journal Rank) - this indicator is a free journal metric inspired by PageRank algorithm,

- SNIP (Source Normalized Impact per Paper) - a factor released in 2012 by Elsevier exactly for Scopus, in order to estimate the citation impact, it weights citations based on the total number of citations in a subject field.

Searching the Scopus database is mostly done over some institution, which pays annually quite large fees for the access. In addition, a single user can register himself after the institutional login was performed, because having a personal account allows the user to save searches and can rerun them in another session, save lists of documents or authors and create document citation alerts and author citation alerts.
Searching for documents itself is advanced by the usage of connectors such as AND, OR, AND NOT.

```
1  PT   AU   BA   BE   GP   AF   BF   CA   TI   SO   SE   BS   LA   DT   CT   CY   CL   SP   HO   DE
   ID   AB   C1   RP   EM   RI   OI   FU   FX   CR   NR   TC   Z9   U1   U2   PU   PI   PA   SN   EI
   BN   J9   JI   PD   PY   VL   IS   PN   SU   SI   MA   BP   EP   AR   DI   D2   EA   EY   PG   WC
   SC   GA   UT   PM   OA   HC   HP   DA
2  J  Pettersen, EF; Goddard, TD; Huang, CC; Couch, GS; Greenblatt, DM; Meng, EC;
   Ferrin, TE          Pettersen, EF; Goddard, TD; Huang, CC; Couch, GS; Greenblatt,
   DM; Meng, EC; Ferrin, TE       UCSF chimera - A visualization system for
   exploratory research and analysis JOURNAL OF COMPUTATIONAL CHEMISTRY
   English Article         molecular graphics; extensibility; visualization;
   multiscale modeling; sequence alignment PROTEIN-STRUCTURE; PECTATE LYASE;
   DATABASE; VIRUS; INFORMATION; STRATEGIES; MICROSCOPY; MODELS; SCOP; CORE  The
   design, implementation, and capabilities of an extensible visualization system,
   UCSF Chimera, are discussed. Chimera is segmented into a core that provides
   basic services and visualization, and extensions that provide most higher level
   functionality. This architecture ensures that the extension mechanism satisfies
   the demands of outside developers who wish to incorporate new features. Two
   unusual extensions are presented: Multiscale, which adds the ability to
   visualize large-scale molecular assemblies such as viral coats, and
   Collaboratory, which allows researchers to share a Chimera session
   interactively despite being at separate locales. Other extensions include
   Multalign Viewer, for showing multiple sequence alignments and associated
   structures; ViewDock, for screening docked ligand orientations; Movie, for
   replaying molecular dynamics trajectories; and Volume Viewer, for display and
   analysis of volumetric data. A discussion of the usage of Chimera in real-world
   situations is given, along with anticipated future directions. Chimera includes
   full user documentation, is free to academic and nonprofit users, and is
   available for Microsoft Windows  Linux  Apple Mac OS X  SGI IRIX  and HP Tru64
```

Figure 3.1: WOS dataset about Information Visualization. [The image was created by the authors of this paper.]

Furthermore, many filters can be applied to refine the search, for example, limiting search results considering date and therefore having documents that have been added to Scopus in the last fourteen days, limiting to document type or access type, etc.

When searching for authors one can refine the search according to the following criteria: source title, affiliation, city, country, subject area. It can also be found how often articles from a particular author have been cited. Search results can afterwards be sorted by one or more different criteria.

In a similar way one can search for affiliations, review documents and author details. There is also a possibility to review the different metrics for a chosen article.

When the simple or even advanced search does not immediately give clear results, one can use Scopus analytical tools to visually analyze and compare the search results.

Figure 3.2 shows a Scopus dataset about Information Visualization. The dataset is a csv file, but can also be downloaded as RIS, bibtex, plain text, and be exported to Mendeley and RefWorks. There are various options to choose which metadata should be exported like citation information, bibliographical information, abstract & keywords, funding details, and other information as shown in Figure 3.3. Many tools allow a direct import from a Scopus csv file. The first row of the csv contains the metadata categories which are separated by commas. Each row contains a document and each column holds the relevant meta data information about this specific document.

Figure 3.2: Scopus dataset about Information Visualization. [The image was created by the authors of this paper.]

Figure 3.3: Scopus export document settings. [The image was created by the authors of this paper.]

Chapter 4

Tools

It is import that the right tools are used for a particular science mapping application. There are a ton of tools out there for constructing and analyzing citation and co-citation networks, bibliographic coupled networks or just general graphs. In this section a concise overview of the most relevant ones in the field is provided. However, the focus in this survey is on dedicated tools.

4.1 General Graph Tools

Co-citation and co-authorship analysis is only a really small part of the whole field of network science. There are dedicated tools to do this kind of analysis, but sometimes a more general tool, which was not intended to be used on science mapping in the first place, is well enough. There are many available on the web.

4.1.1 NodeXL

NodeXL is a free and open source network analysis tool developed by Microsoft under the Microsoft Public Licence. It is available in two versions, namely Basic and Pro, and is an Excel sheet which provides many common metrics on the data.

4.1.2 Gephi

Gephi is an open source network analysis and visualization tool written in Java and developed under the GNU General Public licence [gephi.org 2008]. The ideology is to offer a complementary tool to traditional statistics which enforces visual thinking and helps to intuitively reveal patterns and trends in the graph. Gephi is most often used for exploratory data analysis, which emerged from the field of visual analytics.

4.1.3 GUESS

GUESS [Adar 2019] is a tool, written in Jython and developed under the GNU Public Licence. It supports a fast way to prototype and deploy new visualizations [Adar 2006]. The user easily can add new modules to the GUESS platform by writing Gython code, an extension of Jython. Since Gython is an easy to understand language, full of syntactic sugar, it enables working in an intuitive manner.

4.2 Dedicated Graph Tools

Sometimes a general tool is not enough to analyse the data for co-occurrences or co-citations. The general tools often don't offer a way to import datasets from big knowledge bases like Scopus, Google Scholar or the Web of Science. In this case, a dedicated tool is recommended.
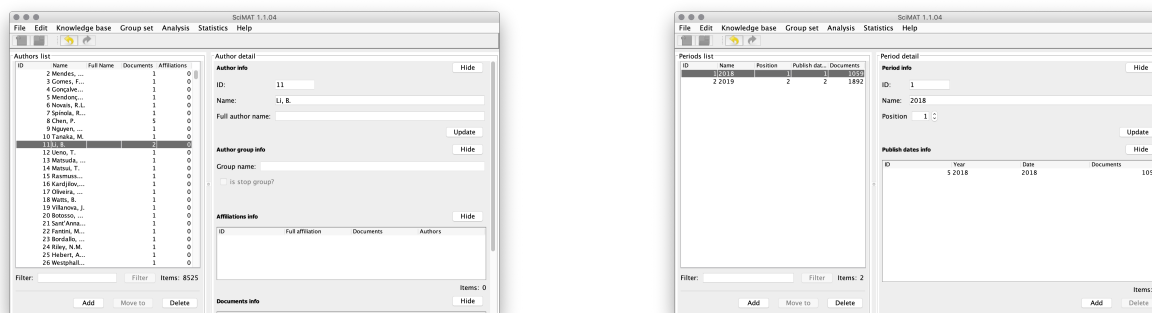
(a) All authors which are currently in the knowledge base of SciMat are displayed here. There is even a possibility to add new ones.

(b) All periods which are currently available in the knowledge base of SciMat.

Figure 4.1: The knowledge base of SciMat. (a) shows the authors tab and (b) the periods tab. [Images are screenshots of the local instance of SciMat on one of the authors machines.]

4.2.1 SciMat

SciMat is an open source software, developed under GPLv3 at the University of Granada. Since it is written in Java, it is cross platform functional as long as your system supports running applications on a Java Virtual Machine. It is divided into three modules.

4.2.1.1 Knowledge Base

The knowledge base provides an interface for adding new data and manipulating it. The user can import new entries from source files in various formats. After the desired number of entries is added to the knowledge base, the data can be reallocated to different periods to handle it easier later. Figure 4.1a shows the knowledge base with all currently stored authors of the dataset and Figure 4.1b the periods in which the user has to separate the dataset to make further analysis.

4.2.1.2 Analysis

SciMat also offers a way to analyse the data in the knowledge base. The user can conduct network analysis to measure for example density and centrality. It is also possible to perform quality analysis to measure values like minimum, maximum or average citations. The analysis component even offers the possibility to do temporal operations.

4.2.1.3 Visualization

After data adding and further modifications and analysis on it is done, the user can visualize the data by utilizing the visualization component of SciMat. However, the specific algorithms for the creation of the graphs must be chosen first. Figure 4.2 shows the wizard which helps the user creating the graph for further inspection. After all variables are set, SciMat needs some time to calculate the graph and to present it afterwards, like shown in Figure 4.3.

4.2.2 VOSviewer

VOSviewer is a tool, free for non-commercial purposes. It was developed in Java at the University of Leiden. It offers beautiful visualizations of any type of bibliometric network with a maximum size of 10.000 citations. Co-citation and bibliographic coupling is supported, but the time dimension is completely ignored.

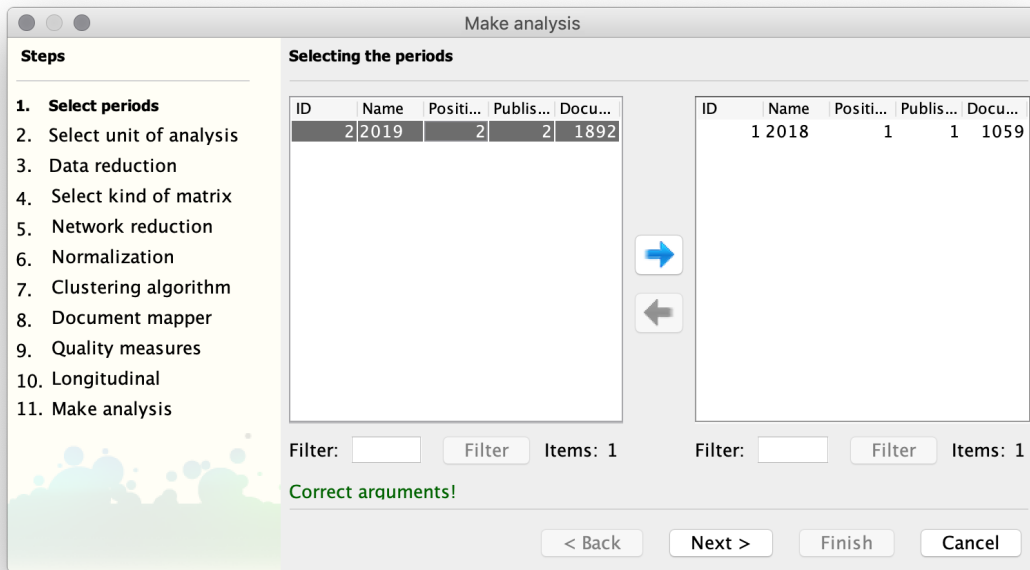**Figure 4.2:** SciMat implements many algorithms into its graph creation component. Before creating a visualization in SciMat, the corresponding algorithms must be chosen. [Image is a screenshot of the local instance of SciMat on one of the authors machines.]
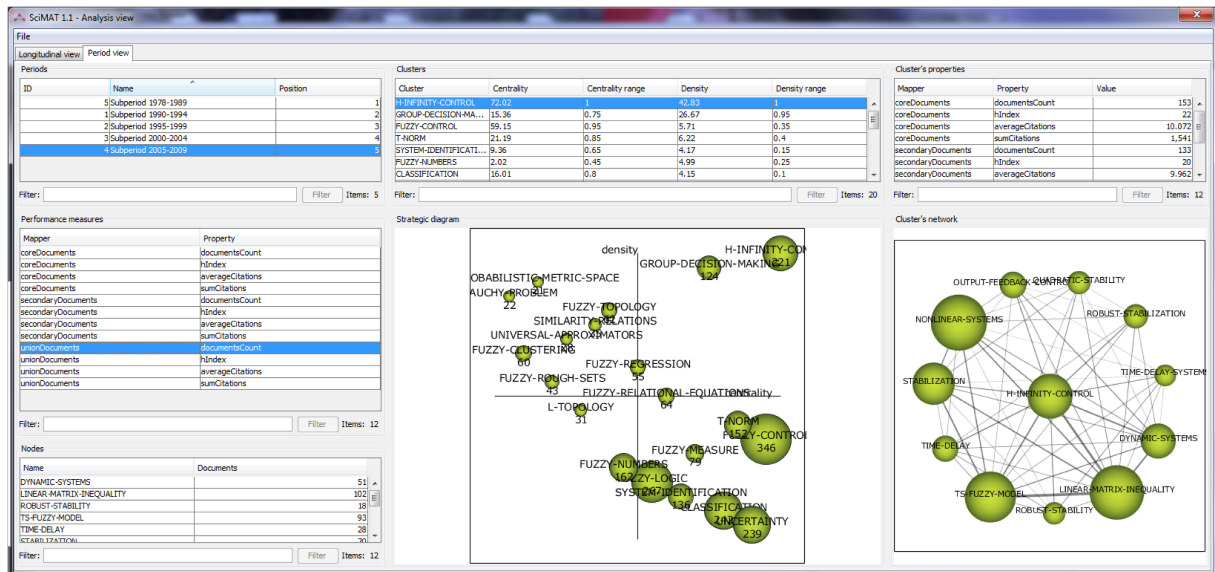
Figure 4.3: SciMat offers many ways to inspect the data of the underlying dataset. [Picture used with kind permission of M. J. Cobo.]
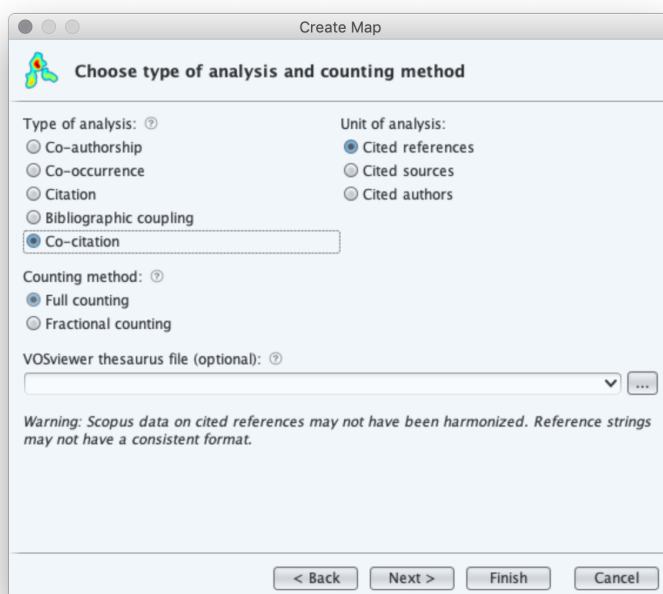
Figure 4.4: VOSviewer provides the option for co-authorship, co-occurrence, citation, bibliographic coupling and co-citation.
[Image is a screenshot of the local instance of VOSviewer on one of the authors machines.]

4.2.2.1 Create a Map

The first thing to do before attempting a VOSviewer visualization, is creating a map based on bibliographic or network data. VOSviewer supports many data formats. Datasets downloaded from Scopus or Web of Science are supported beside other more general formats like CSV or simply plain text. However, the user does not need to create a new map - of course he is free to import an existing map.

4.2.2.2 Science Mapping Options

VOSviewer offers many possibilities how the dataset should be visualized. There is the option for co-authorship, co-occurrence, citation, bibliographic coupling and co-citation. Figure 4.4 shows the window in which the user can inspect and choose the options.

4.2.2.3 Inspecting the Graph

After setting up the map by choosing the mapping options and adjust some settings like isolate outliers, it is now possible to inspect the graph. There are many vertices linked by edges. In this example a co-citation network of information visualization papers was used. A node represents a paper and a link between two nodes means that that those two paper appear together in the reference list of some other paper. Figure 4.5 shows the network.

What attracts attention also is the variation of the node and link size. The bigger the size of a node, the bigger its score in terms of co-citation. The bigger a line between two nodes, the more often they appear together in reference lists. What attracts attention also, are the different colored areas of the network. Every color represents a cluster in the graph or simply nodes which are more related to each other then the rest. This helps researchers to quickly identify papers which are similar or are discussing a similar topic for example. To inspect the different clusters and its containing nodes further, there is an "items" tab. In the items tab all nodes are listed again, providing an option to scroll sideways to read the whole
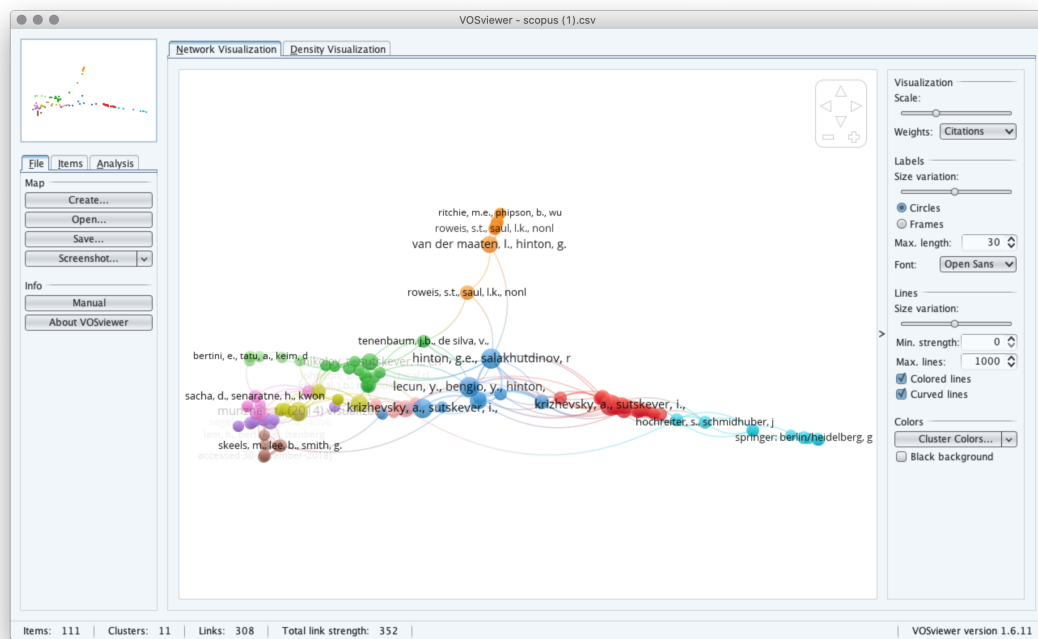
Figure 4.5: After setting up the map by choosing the mapping options and adjusting some settings like isolate outliers, inspection of the graph is possible.
[Image is a screenshot of the local instance of VOSviewer on one of the authors machines.]

paper name. The nodes are sorted by clusters. If a node is double clicked in the tab, VOSviewer instantly zooms into that node in the graph. See the items tab in Figure 4.6a.

4.2.2.4 Analysis

VOSviewer also provides the possibility of analysis on the data. This can be used to update the layout and the clustering of the currently active map [van Eck and Waltman 2019b].

- Normalization of link strength between items
    - **No normalization**: No normalization of link strength.

    - **Association strength**: Normalization based on association strength.

    - **Fractionalization**: Normalization based on fractionalization method.

    - **LinLog/modularity**: Normalization is performed in the same way as in the LinLog layout technique and in the modularity clustering technique.

- Layout
    - **Attraction and repulsion**: Attraction repulsion adjust the location of items in a the map.

    - **Use default values**: Use default values for attraction and repulsion.

- Clustering
    - **Resolution**: Adjust level of detail of clustering.

    - **Min. cluster size**: Defines minimum cluster size.

    - **Merge small clusters**: If minimum cluster size is not met by a cluster, this option defines how small clusters are merged.
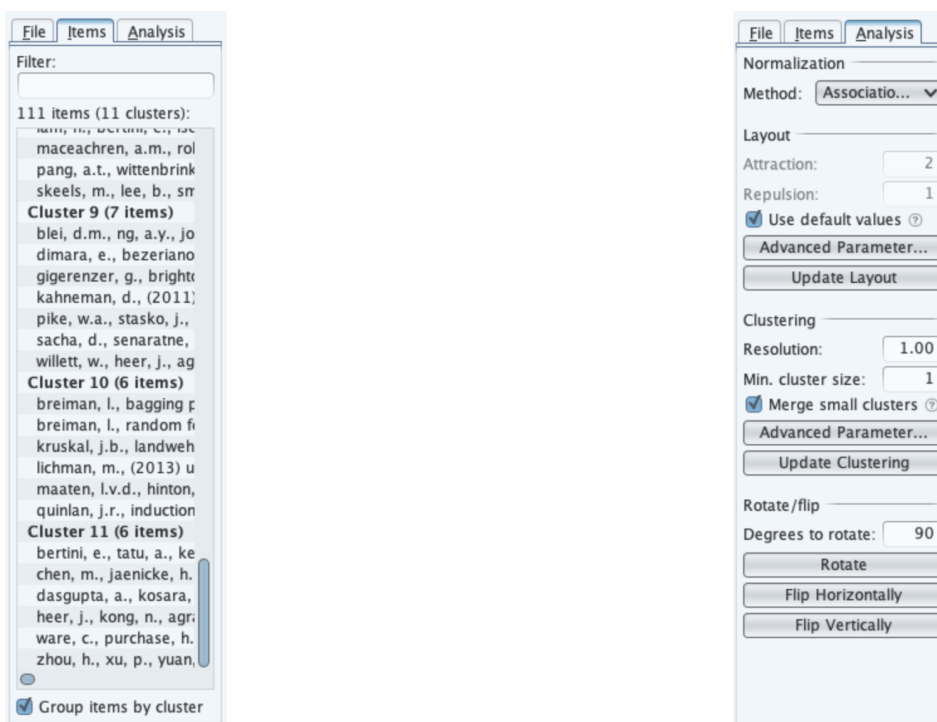
(a) The items tab is providing a compact list-ing of all clusters and their corresponding nodes and an option to scroll sideways to read the whole paper name.

(b) VOSviewer provides the possibility of ana-lysis on the data.

Figure 4.6: The items and analysis tab in VOSviewer. (a) shows the items tab and (b) the analysis tab. [Image is a screenshot of the local instance of VOSviewer on one of the authors machines.]

- Rotate/flip
    - **Rotate**: Rotate the map by a desired number of degrees.
    - **Flip horizontally**: Flip the map horizontally.
    - **Flip vertically**: Flip the map vertically.

The analysis view is listed in Figure 4.6b.

4.2.2.5 Density Network

VOSviewer provides the possibility to view the network in its density representation. This enables the quick spotting of dense parts in the graph. See Figure 4.7

4.2.3 CitNetExplorer

CitNetExplorer [van Eck and Waltman 2019a] is an open-source software written in Java and used to visualize and analyze citation networks of publications. It is free to use for non-commercial purposes and was developed at the Leiden University by the same developers who also developed VOSviewer. CitNetExplorer allows the direct import of a downloaded dataset from the Web of Science database and supports very large citation networks.
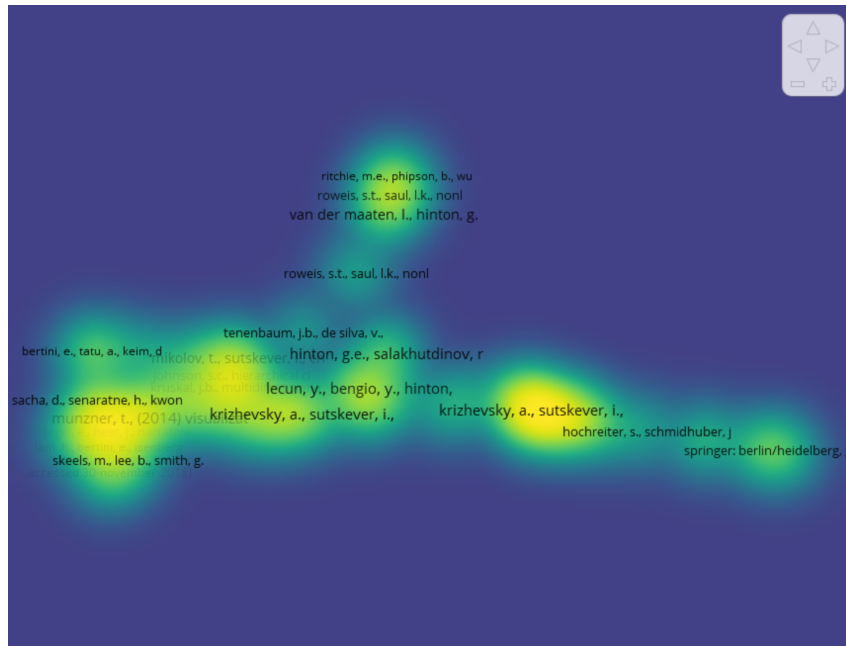
Figure 4.7: VOSviewer provides the possibility of a density visualization. [Image is a screenshot of the local instance of VOSviewer on one of the authors machines.]

(a) Publications displayed as nodes.
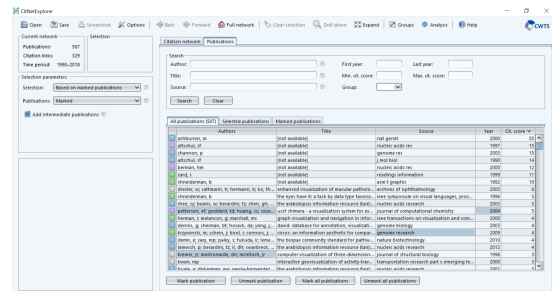
(b) Various information available of publications including the citation score.

Figure 4.8: Interface of CitNetExplorer. There are two tabs at the top of the main view from which the user can switch between. (a) shows the citation network and (b) the publications. [The screenshots were created with CitNetExplorer by the authors of this paper.]

4.2.3.1 Visualization

The main overview provides two tabs from where the user can interactively switch between citation network view and publications as shown in Figure 4.8. In the citation network view the nodes represent publications labeled by the last name of the first or last author and the edges represent direct or indirect citation relations. Direct citation relations have darker lines whereas indirect relations have lighter lines. The vertical axis represents time which means nodes at the top are old publications and nodes at the bottom are more recent. The horizontal axis represents the citation relatedness. The closer the nodes are to each other the stronger is their citation relation and the further away the nodes are the weaker is their relation. The direct and indirect relation is taken into account. The cited publication is always above the citing publication.

In the publication tab the user can search for author, title, source, first or last year, min or max citation score, and group. By right clicking on one publication the user can assign this publication to a different
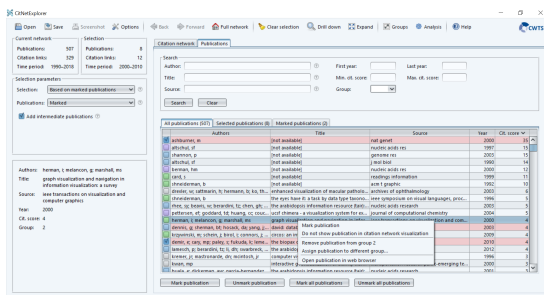
(a) The publication can be viewed in the web browser.
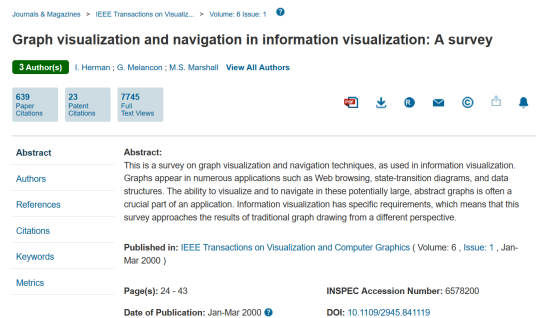
(b) The publication will be displayed in a web browser with all provided information where it can be read and downloaded.

Figure 4.9: Publication options in CitNetExplorer. (a) shows the available options in CitNetExplorer of a publication and (b) shows the publication in a web browser. [The screenshots were created with CitNetExplorer by the authors of this paper.]

(a) The search function lets the user narrow down the publications.
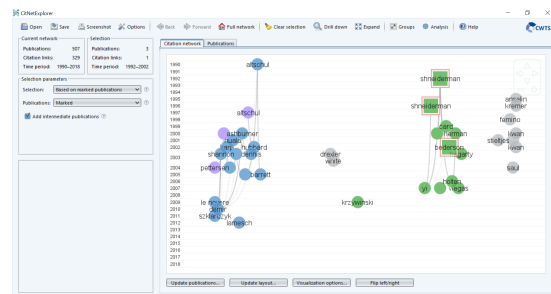
(b) Selecting the displayed publications from the previous search result leads also to a selection in the graph.

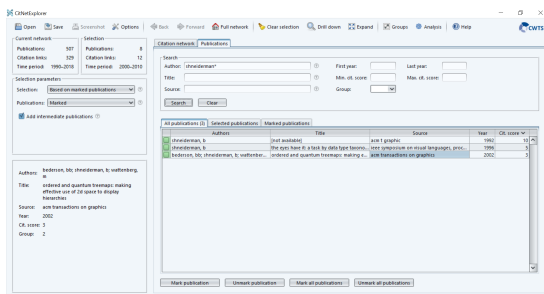**Figure 4.10:** Search function in CitNetExplorer. (a) shows the search results for the author "shneiderman" and (b) shows the selected publications in the graph. [The screenshots were created with CitNetExplorer by the authors of this paper.]

group or open the publication in the web browser as shown in Figure 4.9. When the user performs a search the results will be displayed at the bottom where the user can also mark or unmark all displayed publications. The marked publications will be highlighted in the citation network view by a square around the node and a light red color. This process is displayed in Figure 4.10.

4.2.3.2 Analysis

There are three different ways to analyze a citation network.

1. **Selection of publications:** In the selection parameters window panel the user can choose between three options: Based on marked publications, time period, and groups as shown in Figure 4.11a. The default setting is based on marked publications. In general, the user can select multiple individuals per mouse and they appear as marked in the main panel or mark multiple documents in the publication tab. For time period the user can choose a first and last year and all publications between the two years are going to be selected. The last one is based on groups which displays all current groups. The user can choose which group should be displayed in the main panel by selecting one or multiple groups and then with the help of the drill down feature to narrow down the publications.

2. **Drill down and expand:** The drill down feature offers the user a possibility to only display the
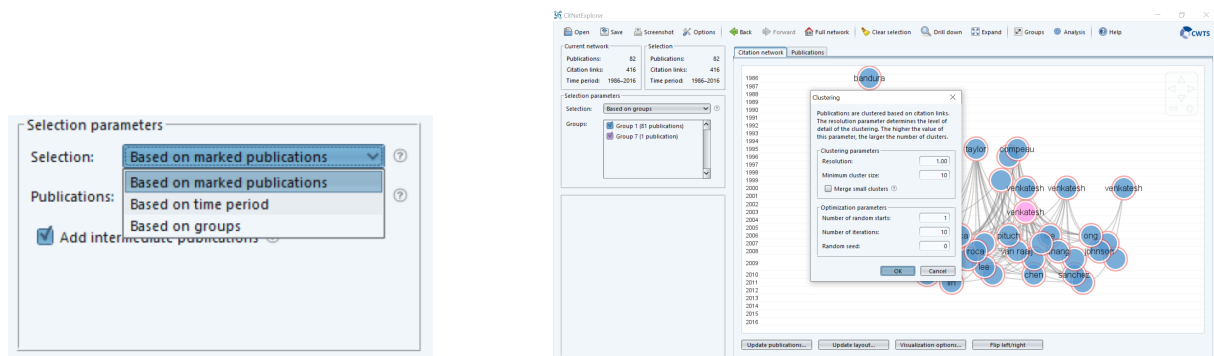
(a) There are three different selection options available.

(b) The default settings for clustering.

Figure 4.11: Some of the analysis options in CitNetExplorer. (a) shows the selection panel and (b) the clustering panel. [The screenshots were created with CitNetExplorer by the authors of this paper.]

selected or marked publications in the citation network window, whereas the expand feature offers the user to show the predecessors and successors of the publications. Additionally, a back and forward button are provided which have a similar functionality to the ones in a web browser, with which the user can go back and forth between actions.

3. **Algorithms:** There are various algorithms available like identifying the largest or all connected components, core publications, shortest and longest paths or perform clustering. For each of these algorithms the user has additional adjustments available but can also choose the provided default settings as shown in Figure 4.11b for clustering.

4.2.4 Sci2

The Science of Science (Sci2) tool [Sci2 Team 2009] is designed for the study of science and supports the temporal, geospatial, topical, and network analysis and visualization of various datasets. Sci2 was developed by the Sci2 Team at the University of Indiana and is build upon the Cyberinfrastructure Shell (CIShell), which is an open source software framework written in Java. CIShell simplifies the integration of datasets, algorithms, tools, and computing resources and greatly facilitates the utilization for the user.

The user can load their own datasets or access the provided datasets in the "sampledata" folder. Various types of analysis can be performed due to the highly efficient algorithms which are available. Sci2 allows the user to use different visualizations based on the desired outcome and lets the user explore, research, and understand datasets and gather useful information about it. The functionality of Sci2 consists of:

- Loading Data (Files)

- Data Preparation

- Preprocessing

- Analysis

- Modeling

- R

- Visualization

4.2.4.1 Functionality

The focus of this work will be on network analysis and visualization. This paragraph provides a short overview about the functionality of Sci2 and further information can be found at [Sci2 Team 2011]. After the dataset has been imported Sci2 provides several network extraction methods:

- **Extract Directed Network:** Directed network is created by defining a directed edge between the values of two columns.

- **Extract Bipartite Network:** Unweighted bipartite network is created by defining a directed edge between the values of two columns.

- **Extract Paper Citation Network:** Unweighted directed network is extracted from papers and their corresponding citations.

- **Extract Author Paper Network:** Unweighted directed network is extracted from authors and their corresponding papers.

- **Extract Co-Occurrence Network:** Network is extracted from a delimited table.

- **Extract Word Co-Occurrence Network:** Weighted network is created by defining edges between words whereas a node represents a word. The weight of an edge is defined on how often two words appear together in the same text.

- **Extract Co-Author Network:** Weighted network is extracted from authors which appear together in a paper. A node represents an author and the weight of an edge is defined on how often two authors wrote a paper together.

- **Extract Reference Co-Occurrence (Bibliographic Coupling) Network:** Weighted network is extracted from a paper citation network where papers are nodes and the weight of an edge is defined on how many citations two papers have in common.

- **Extract Document Co-Citation Network:** Weighted network is extracted from a paper citation network where papers are nodes and the weight of an edge is defined on how often two papers have been cited together.

There are several options available to clean the data and to perform advanced algorithms for analysis, some frequent used are listed below the rest can be seen at [Lind and Kloster 2015]:

Preprocessing:
- Extract Top Nodes
- Extract Nodes Above or Below Value
- Delete Isolates
- Extract Top Edges
- Extract Edges Above or Below Value
- Remove Self Loops
- and many more

Analysis:
- Node Degree
- Degree Distribution
- K-Nearest Neighbor
- Clustering Coefficient
- HITS
- PageRank
- and many more

There are two main integrated tools to visualize network data which are GUESS and Gephi. GUESS is a free database driven system which is written in Java, open-source, and contains a console that allows the user to script in Jython (lets the user code in Python and runs on a Java platform). The tool is specialized in data analysis, exploring, and visualizing graphs and networks. However, there are some limitations when it comes to dealing with large networks which have more than 20.000 nodes. Gephi is an
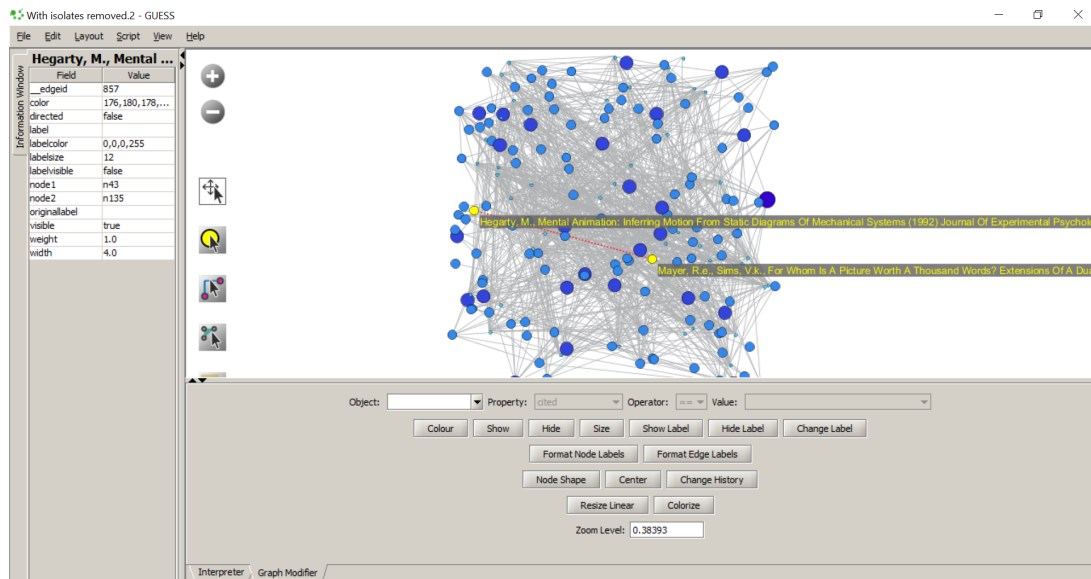
Figure 4.12: Co-citation network created in Sci2 and visualized with GUESS. [The image is created with Sci2 by the authors of this paper.]

open-source software and free. It allows the user to interactively explore and visualize networks, systems, and more. Supported are GraphML files or any network files which can be converted to GraphML [Adar 2007; gephi.org 2008].

4.2.4.2 Co-Citation Network

To create a co-citation network a dataset will be needed first. The dataset is from Scopus about information visualization, has 2.000 entries and contains different document types such as conference papers, articles, reviews, conference reviews, book chapters, articles in press, books, surveys, editorials, and notes. Next a paper citation network needs to be extracted from the imported dataset in order to extract the document co-citation network. This process takes a few minutes and after the co-citation network has been extracted the network can be cleaned. Isolated nodes and self loops are deleted and the top 200 nodes based on the citation count are extracted since GUESS does not handle large networks very well. Figure 4.12 shows the co-citation network in GUESS and Figure 4.13 in Gephi. The nodes have been resized according to their citation count, the higher the citation count the larger the node will be. The node color has been adapted to the citation count too. Nodes with a darker color have a higher citation count and nodes with a lighter color have a lower citation count.

4.2.4.3 Co-Authorship Network

The same dataset was used to extract the co-authorship network. Preprocessing the data leads to deleting the isolates and extracting the top 200 nodes based on the number of authored works. The nodes are resized and recolored according to the number of authored works and the result is shown in Figure 4.14 and Figure 4.15.

4.2.4.4 Bibliographic Coupling

To create a co-occurrence (bibliographic coupling) network first a paper citation network needs to be extracted from the imported dataset. The cleaned and further preprocessed network is shown in Figure 4.16 and Figure 4.17. The edges are resized and recolored based on their weights, which means the thickness of the edge represents how often two papers have been cited together.

Figure 4.13: Co-citation network created in Sci2 and visualized with Gephi. [The image is created with Sci2 by the authors of this paper.]

Figure 4.14: Co-authorship network created in Sci2 and visualized with GUESS. [The image is created with Sci2 by the authors of this paper.]

Figure 4.15: Co-authorship network created in Sci2 and visualized with Gephi. [The image is created with Sci2 by the authors of this paper.]

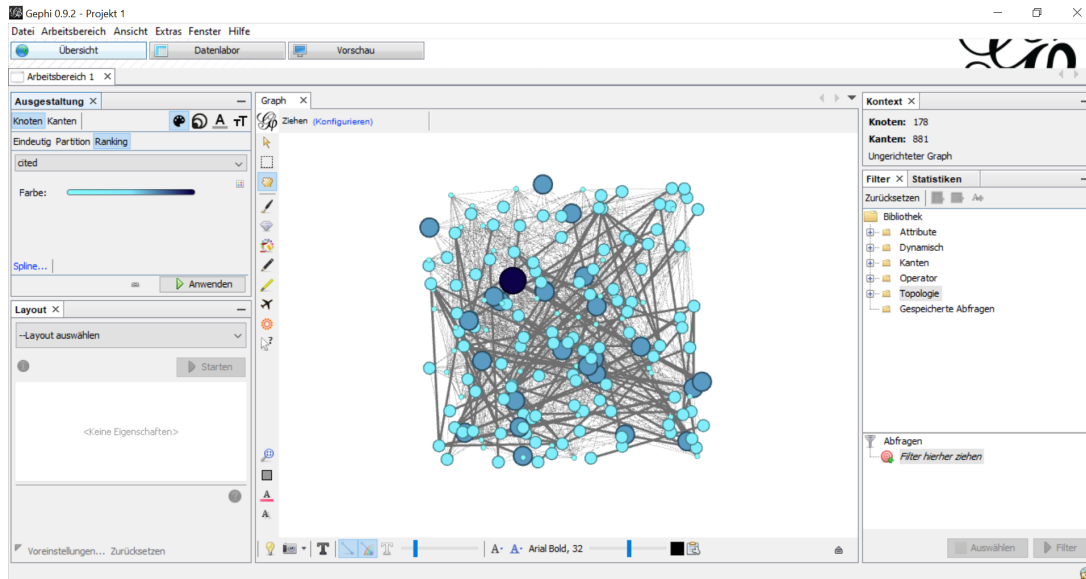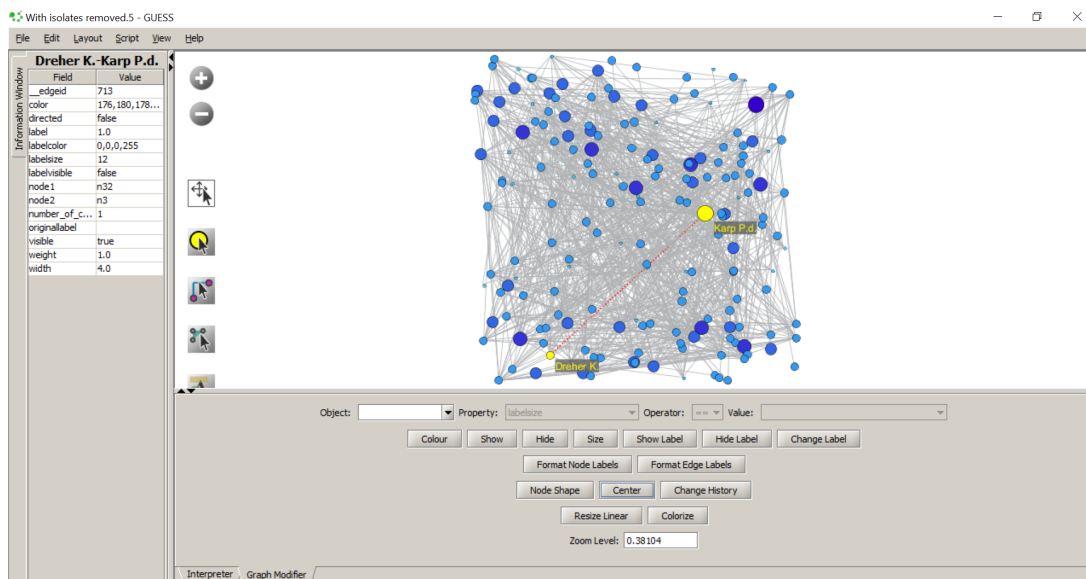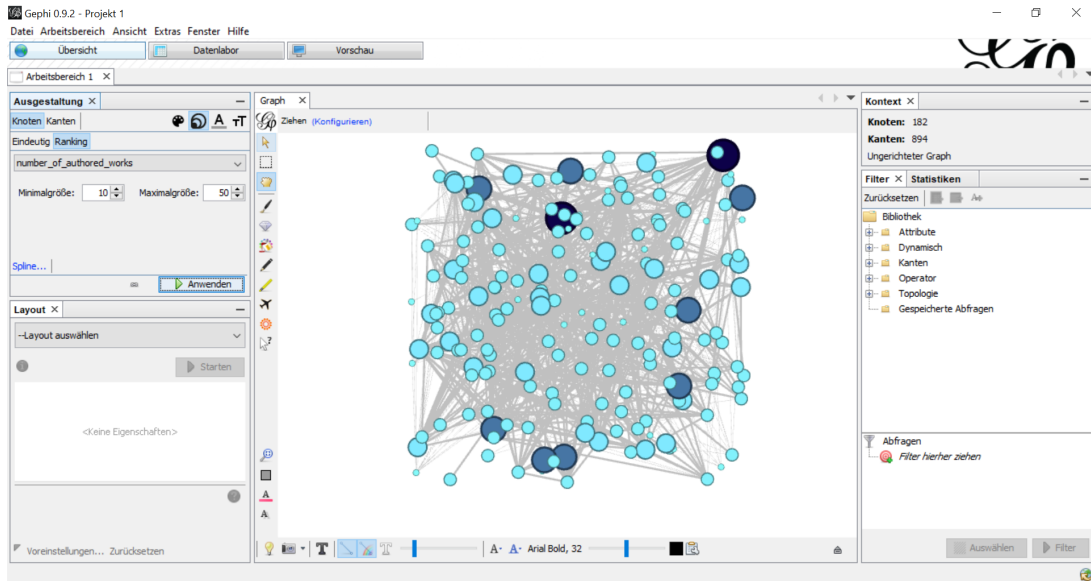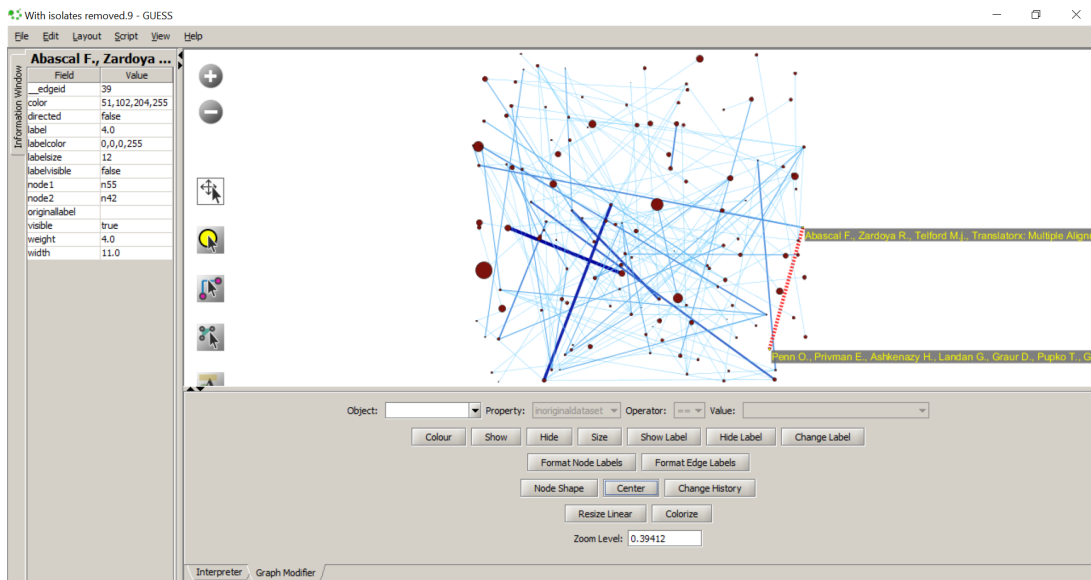**Figure 4.16:** Bibliographic coupling network created in Sci2 and visualized with GUESS. [The image is created with Sci2 by the authors of this paper.]
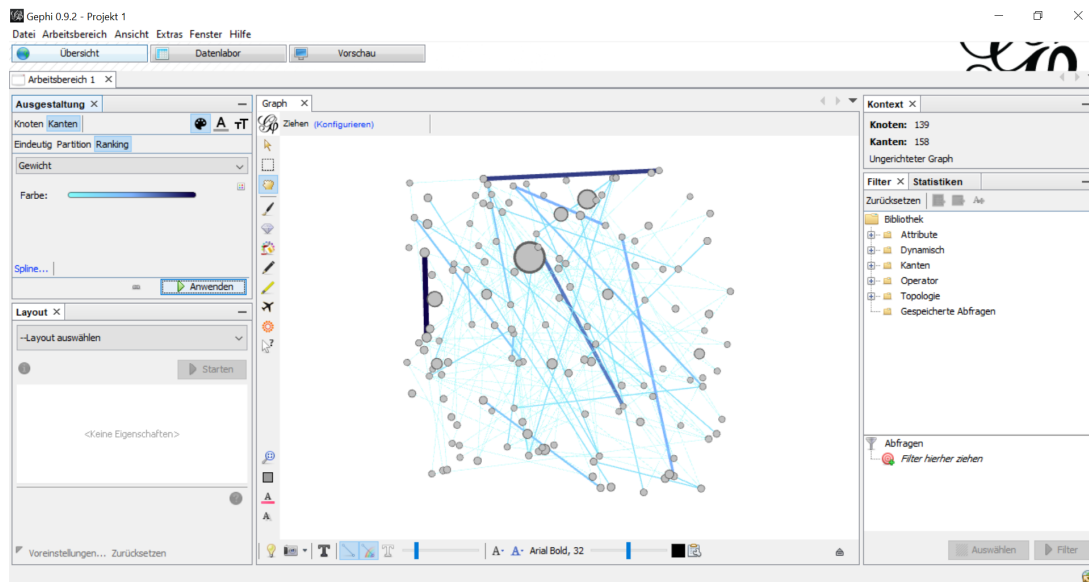
Figure 4.17: Bibliographic coupling network created in Sci2 and visualized with Gephi. [The image is created with Sci2 by the authors of this paper.]

Chapter 5

Miscellaneous

5.1 Places & Spaces: Mapping Science

Places & Spaces: Mapping Science [Cyberinfrastructure for Network Science Center 2019] is an exhibition where an international advisory board chooses annually one outstanding submission of how scientific data can be specially visualized. The exhibition has been displayed at over 375 venues all around the world and is curated by the Cyberinfrastructure for Network Science Center which is located at Indiana University. There are four types of exhibition setups: The Physical Exhibit, Maps-Only Physical Exhibit, Poster Exhibit, and Digital Exhibit. At least once a year, ten new maps are added to the collection of Places & Spaces which can be viewed in the online gallery.

In 2005 Klavans and K. Boyack [2006] presented the most accurate literature map about research communities which covered hundreds of thousands of scientific papers. Later in 2007 they provided a science map forecasting how the structure of science may change in the future. With the help of the Mercator projection the map remained two dimensional and showing an increase or decrease of the connectedness between various scientific fields. Both maps are visualized in [Börner 2010].

In 2017 one entry was a web application by Bornmann et al. [2016b] which shows the collaboration of the top research institutions and universities. Each institution was categorized in subject areas and a co-authorship network shows with which institutions other institutions have collaborated and how successfully. The interactive interface lets the user choose a subject area and the layout can be switched between network and geographic. Individual institutes can be clicked on, displaying further information and interaction. The map is automatically zoomed in by clicking on an institution listed at the bottom, but the user can turn this feature off. A colorization can be selected based on the country or citation count of papers. The application can be found at [Bornmann et al. 2016a] and is shown in Figure 5.1.

Another entry in 2017 was "Megaregions of the US" by Nelson and Alasdair [2016] who displayed a map of the United States which shows the linkage of commuters between one area to another taken on the census data from 2010. To identify any regional patterns the community detection algorithm was used to extract groups and form clusters of significant locations. Nelson and Alasdair [2016] talk about how difficult it is to assign commutes to a single community that cross different communities and resulted in using a computational-visual approach. Convex hulls were traced around communities, boundary lines cleaned up, outliers eliminated and geographic contiguities highlighted. The result were four million lines which look like starbursts as shown in Figure 5.2.
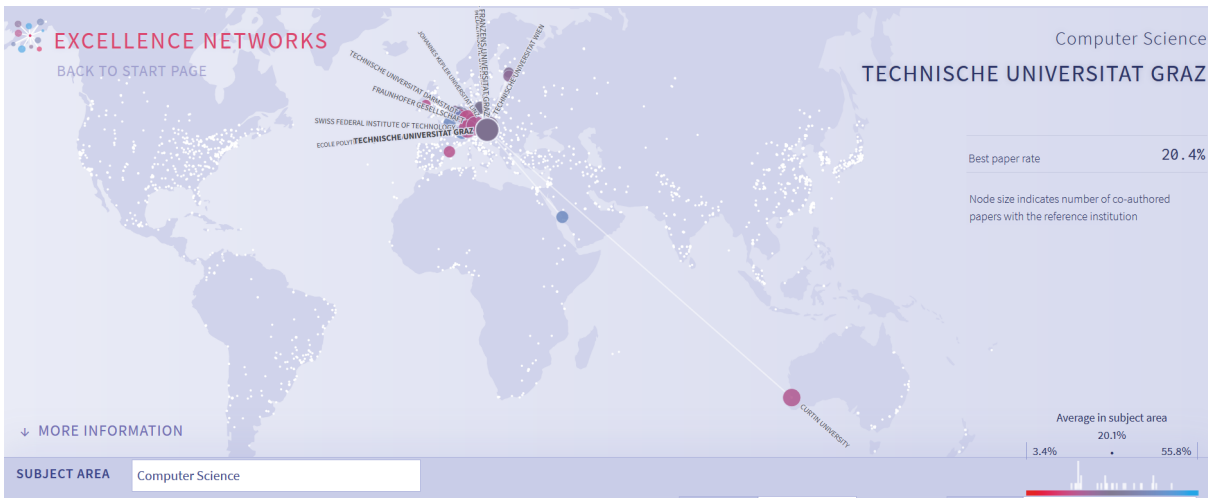
Figure 5.1: Excellence network showing the Technical University of Graz and their collaboration with others in the subject area of computer science. The papers have been published between 2011 and 2015. [Screenshot taken from the excellence networks web application, original at [Bornmann et al. 2016a] used with kind permission of M. Stefaner, L. Bornmann, R. Mutz, and F. M. Anegon.]
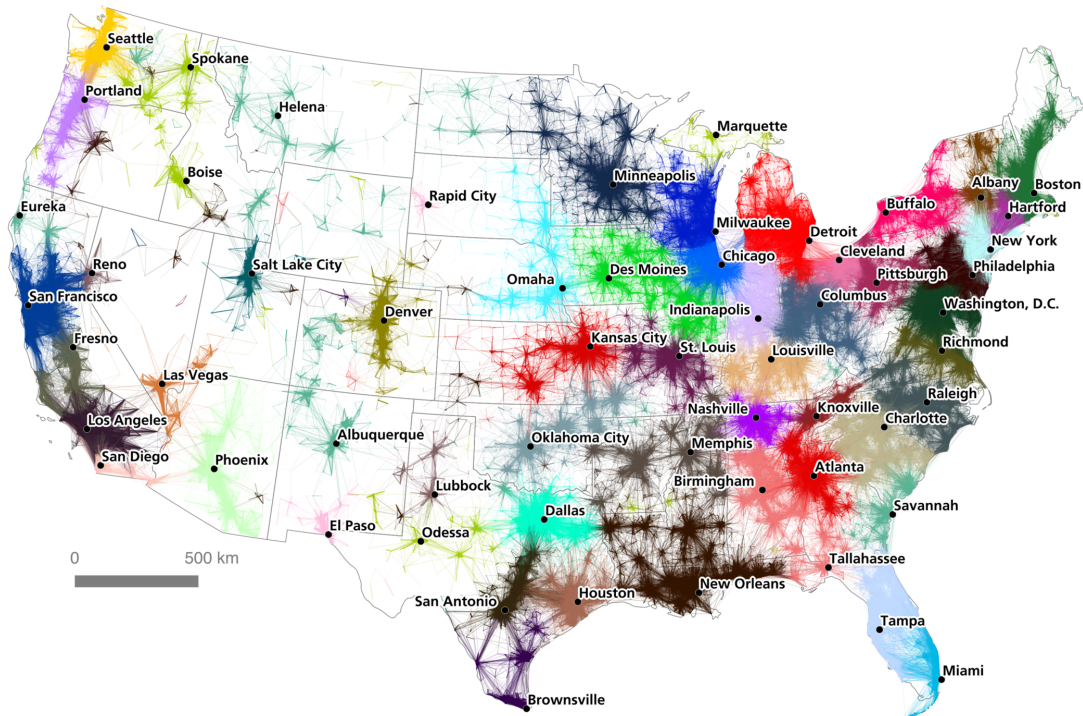
Figure 5.2: A commuter flow-based regionalization of the United States. [Image extracted from [Nelson and Alasdair 2016] and used under the terms of the CC-BY 4.0. Copyright © by the Creative Commons Attribution 4.0 License]

Chapter 6

Concluding Remarks

The ideas behind science mapping, its use in the scientific community and its purpose and implications were discussed. An introduction to the most common applied methods and tools in the field was given, including a short overview of two largest citation databases. Using the right tool is not always an obvious decision. However, if the time dimension is important then Sci2 or CitNetExplorer should be used. Sci2 provides also basic graph properties for further analysis, but does not handle large datasets very well. The memory can be be redefined, but the correct balance between choosing enough memory and preventing to crash the tool has to be found. CitNetExplorer is probably the way to go if massive datasets should be supported and performance is not an issue. If a more hierarchical structure of the graph is preferred CitNetExplorer is also a good option since it displays publications based on the time and can hide irrelevant publications. However, if performance is an issue, then the recommendation is VOSviewer since it performs very fast. SciMat provides by far the most clustering algorithms and should be used if that is a necessity.

Bibliography

Adar, Eytan [2006]. *GUESS: A Language and Interface for Graph Exploration*. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '06. ACM, 2006, pages 791–800. ISBN 1-59593-372-7. doi:10.1145/1124772.1124889. doi.acm.org/10.1145/1124772.1124889 (cited on page 11).

Adar, Eytan [2007]. *GUESS: The Graph Exploration System*. 13 Aug 2007. graphexploration.cond.org (cited on page 21).

Adar, Eytan [2019]. *GUESS Website*. graphexploration.cond.org/index.html (cited on page 11).

Börner, Katy [2010]. *Atlas of Science: Visualizing what We Know*. MIT Press, 17 Sep 2010, page 272. ISBN 0262014459 (cited on page 25).

Bornmann, Lutz, Rüdiger Mutz, Moritz Stefaner, and Felix de Moya Anegon [2016a]. *Excellence Networks*. Excellence Networks, 2016. excellence%C2%AD-networks.net (cited on pages 25–26).

Bornmann, Lutz, Rüdiger Mutz, Moritz Stefaner, and Felix de Moya Anegon [2016b]. *Excellence networks in science: A Web-based application based on Bayesian multilevel logistic regression (BMLR) for the identification of institutions collaborating successfully*. Journal of Informetrics 10.1 (2016), pages 312–327. ISSN 1751-1577. doi:10.1016/j.joi.2016.01.005. arxiv.org/abs/1508.03950 (cited on page 25).

Boyack, Kevin W., Richard Klavans, and Katy Börner [2005]. *Mapping the Backbone of Science*. Scientometrics 64.3 (Aug 2005), pages 351–374. doi:10.1007/s11192-005-0255-6 (cited on page 1).

Chen, Chaomei, Rachael Dubin, and Timothy Schultz [2017]. *Science Mapping*. In: Jul 2017, pages 271–284. ISBN 1466658886. doi:10.4018/978-1-4666-5888-2.ch410 (cited on page 1).

Clarivate [2018]. *Web of Science Core Collection Field Tags*. Clarivate, 09 Oct 2018. images.webofknowledge.com/images/help/WOS/hs_wos_fieldtags.html (cited on page 8).

Clarivate Analytics [2019a]. *Clarivate Analytics*. clarivate.com (cited on page 7).

Clarivate Analytics [2019b]. *Web of Science*. webofknowledge.com (cited on page 7).

Clarivate Analytics [2019c]. *Web of Science Core COllection*. clarivate.com/products/web-of-science/databases (cited on page 7).

Colavizza, Giovanni, Kevin W. Boyack, Nees Jan van Eck, and Ludo Waltman [2018]. *The Closer the Better: Similarity of publication pairs at different co-citation levels*. Journal of the Association for Information Science and Technology 69 (2018), pages 600–609. doi:10.1002/asi.23981. http://onlinelibrary.wiley.com/doi/10.1002/asi.23981 (cited on page 3).

Cyberinfrastructure for Network Science Center [2019]. *Places & Spaces: Mapping Science*. Cyberinfrastructure for Network Science Center, 2019. scimaps.org (cited on page 25).

Egghe, Leo and Ronald Rousseau [2002]. *Co-citation, bibliographic coupling and a characterization of lattice citation networks*. Scientometrics 55 (Nov 2002). doi:10.1023/A:1020458612014. researchgate.net/

publication/228815594_Co-citation_bibliographic_coupling_and_a_characterization_of_lattice_
citation_networks (cited on pages 3–5).

Elsevier [2019a]. *Elsevier*. elsevier.com (cited on page 8).

Elsevier [2019b]. *Scopus*. scopus.com (cited on page 8).

gephi.org [2008]. *Gephi: The Open Graph Viz Platform*. gephi.org, 2008. gephi.org (cited on pages 11, 21).

Kessler, M. M. [1963]. *Bibliographic coupling between scientific papers*. American Documentation 14.1 (1963), pages 10–25. doi:10.1002/asi.5090140103. eprint: onlinelibrary.wiley.com/doi/pdf/10.1002/ asi.5090140103. onlinelibrary.wiley.com/doi/abs/10.1002/asi.5090140103 (cited on page 3).

Klavans, Richard and Kevin Boyack [2006]. *Quantitative evaluation of large maps of science*. Scientometrics 68 (Dec 2006), pages 475–499. doi:10.1007/s11192-006-0125-x (cited on page 25).

Lind, Sean and David Kloster [2015]. *Sci2 Algorithms and Tools*. Indiana University and SciTech Strategies, 17 Aug 2015. wiki.cns.iu.edu/display/SCI2TUTORIAL/3.1+Sci2+Algorithms+and+Tools#id-3.1Sci2AlgorithmsandTools-Networks (cited on page 20).

McCain, Katherine W. [1991]. *Mapping Economics through the Journal Literature: An Experiment in Journal Cocitation Analysis*. Journal of the American Society for Information Science 32 (May 1991), pages 290–296. doi:10.1002/(SICI)1097-4571(199105)42:4<290::AID-ASI5>3.0.CO;2-9 (cited on page 4).

Nelson, Garrett Dash and Rae Alasdair [2016]. *An Economic Geography of the United States: From Commutes to Megaregions*. PLOS ONE 11.11 (Nov 2016), pages 1–23. doi:10.1371/journal.pone.0166083. doi.org/10.1371/journal.pone.0166083 (cited on pages 25–26).

Nguyen Tat Thang [2019]. *List of Science Citation Index Expanded journals 2019*. May 2019. researchgate.net/publication/331938969_List_of_Science_Citation_Index_Expanded_journals_2019_Update_this_month (cited on page 7).

Ponomariov, Branco and Craig Boardman [2016]. *What is co-authorship?* Scientometrics 109.3 (2016), pages 1939–1963. ISSN 1588-2861. doi:10.1007/s11192-016-2127-7. doi.org/10.1007/s11192-016-2127-7 (cited on page 3).

Sci2 Team [2009]. *Sci2 Tool : A Tool for Science of Science Research and Practice*. Indiana University and SciTech Strategies, 2009. sci2.cns.iu.edu (cited on page 19).

Sci2 Team [2011]. *Science of Science (Sci2) Tool Manual*. Indiana University and SciTech Strategies, 28 Mar 2011. wiki.cns.iu.edu/display/SCI2TUTORIAL (cited on page 20).

Small, Henry [2018]. *Citation Indexing Revisited: Garfield's Early Vision and Its Implications for the Future*. Frontiers in Research Metrics and Analysis (Mar 2018). doi:10.3389/frma.2018.00008. frontiersin.org/articles/10.3389/frma.2018.00008 (cited on page 7).

Van Eck, Nees Jan and Ludo Waltman [2019a]. *CitNetExplorer - Analyzing citation patterns in scientific literature*. Centre for Science and Technology Studies, 2019. citnetexplorer.nl (cited on page 16).

Van Eck, Nees Jan and Ludo Waltman [2019b]. *VOSviewer Manual*. Leiden University, 01 Apr 2019. vosviewer.com/documentation/Manual_VOSviewer_1.6.10.pdf (cited on page 15).

White, Howard and Belver C. Griffith [1981]. *Author Cocitation: A Literature Measure of Intellectual Structure*. Journal of the American Society for Information Science 32 (May 1981), pages 163–171. doi:10.1002/asi.4630320302 (cited on page 4).