Chapter 14

Replicable Science of Science Studies

Katy Börner and David E. Polley

Abstract Much research in bibliometrics and scientometrics is conducted using proprietary datasets and tools making it hard if not impossible to replicate results. This chapter reviews free tools, software libraries, and online services that support science of science studies using common data formats. We then introduce plug-and-play macroscopes (Börner, 2011) that use the OSGi industry standard to support modular software design, i.e., the plug-and-play of different data readers, preprocessing and analysis algorithms, but also visualization algorithms and tools. Exemplarily, we demonstrate how the open source Science of Science (Sci2) Tool can be used to answer temporal (when), geospatial (where), topical (what), and network questions (with whom) at different levels of analysis—from micro to macro. Using the Sci2 Tool, we provide hands-on instructions on how to run burst analysis (Song & Chambers, 2014), overlay data on geospatial maps (Hardeman, Hoekman, & Frenken, 2014), generate science map overlays, and calculate diverse network properties, e.g., weighted PageRank (Waltman & Yan, 2014) or community detection (Milojevic, 2014), using data from Scopus, Web of Science or personal bibliography files, e.g., EndNote or BibTex. We exemplify tool usage by studying evolving research trajectories of a group of physicists over temporal, geospatial, and topic space as well as their evolving co-author networks. Last but not least, we show how plug-and-play macroscopes can be used to create bridges between existing tools, e.g., Sci2 and the VOSviewer clustering algorithm (Van Eck & Waltman, 2014), so that they can be combined to execute more advanced analysis and visualization workflows.

14.1 Open Tools for Science of Science Studies

Science of science studies seek to develop theoretical and empirical models of the scientific enterprise. Examples include qualitative and quantitative methods to estimate the impact of science (Cronin & Sugimoto, 2014) or models to understand the production of science (Scharnhorst, Börner, & van den Besselaar, 2012). There exist a variety of open-source tools that support different types of analysis and visualization. Typically, tools focus on a specific type of analysis and perform well at a certain level, e.g., at the micro (individual) level or meso level—using datasets containing several thousand records. What follows is a brief survey of some of the more commonly used tools that support four types of analyses and visualizations: temporal, geospatial, topical, and network.

Many tools have a temporal component, but few are dedicated solely to the interactive exploration of time-series data. The Time Searcher project is an excellent example of a tool that allows for interactive querying of time stamped data through the use of timeboxes, a graphical interface that allows users to build and manipulate queries (Hochheiser & Shneiderman, 2004). Now in its third iteration, the tool can handle more than 10,000

Katy Börner
Cyberinfrastructure for Network Science Center
School of Informatics and Computing, Indiana University
1320 E. Tenth Street, Bloomington, IN 47405, USA

David E. Polley
University Library
Indiana University-Purdue University Indianapolis
755 West Michigan Street, Indianapolis, IN 46292, USA

data points and offers data-driven forecasting through a Similarity-Based Forecasting (SBF) interface (Buono et al., 2007). The tool runs on Windows platforms and is freely available.

There are many tools available for advanced geospatial visualization and most require some knowledge of geographical information science. Exemplary tools include GeoDa, GeoVISTA, and CommonGIS. GeoDa is an open-source and cross-platform tool that facilitates common geospatial analysis functionality, such as spatial autocorrelation statistics, spatial regression functionality, full space-time data support, cartograms, and conditional plots (and maps) (Anselin, Syabri, & Kho, 2006). Another tool, GeoVISTA Studio, comes from the GeoVISTA Center at Penn State, which produces a variety of geospatial analysis tools. Studio is an open-source graphical interface that allows users to build applications for geocomputation and visualization and allows for interactive querying, 3D rendering of complex graphics, and 2D mapping and statistical tools (Takatsuka & Gahegan, 2002). Finally, CommonGIS is a java-based geospatial analysis tool accessible via the web. This service allows for interactive exploration and analysis of geographically referenced statistical data through any web browser (Andrienko et al., 2002).

There are a variety of topical data analysis and visualization tools used in a variety of disciplines from bibliometrics to digital humanities and business. TexTrend is a freely available, cross-platform tool that aims to support decision-making in government and business. Specifically, the tool facilitates text mining and social network analysis, with an emphasis on dynamic information (Kampis, Gulyas, Szaszi, Szakolczi, & Soos, 2009). The aim of the tool is to extract trends and support predictions based on textual data. VOSviewer is another cross-platform and freely available topical analysis tool, designed specifically for bibliometric analysis (Van Eck & Waltman, 2014). The tool allows users to create maps of publications, authors, and journals based on co-citation networks or keywords, illuminating the topic coverage of a dataset (Van Eck & Waltman, 2014). VOSviewer provides multiple ways to visualize data, including a label view, density view, cluster density view, and a scatter view.

Finally, there are many tools that are dedicated to network analysis and visualization. Some of the more prominent tools include Pajek and Gephi. Pajek has long been popular among social scientists. Originally designed for social network analysis, the program is not open-source but is freely available for non-commercial use on the Windows platform. Pajek is adept at handling large networks containing thousands of nodes (Nooy, Mrvar, & Batageli, 2011). The tool handles a variety of data objects, including networks, partitions, permutations, clusters, and hierarchies. Pajek offers a variety of network analysis and visualization algorithms and includes bridges to other programs, such as the ability to export to R for further analysis. Gephi is a more recent, and widely used network analysis and visualization program (Bastian, Heymann, & Jacomy, 2009). This tool is open-source and available for Windows, Mac OS, and Linux. Gephi is capable of handling large networks, and provides common network analysis algorithms, such as average network degree, graph density, and modularity. Gephi is also well suited for displaying dynamic and evolving networks and accepts a wide variety of input formats including NET and GRAPHML files.

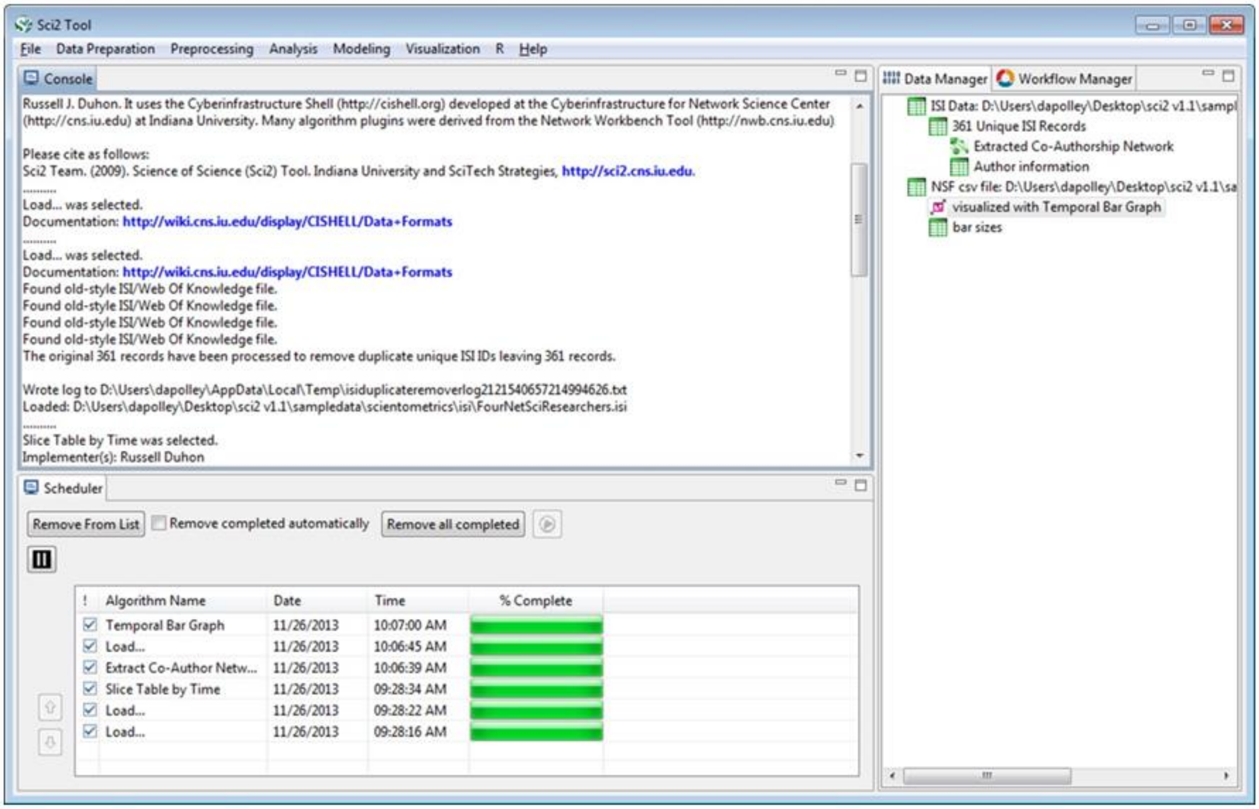
14.2 The Science of Science (Sci2) Tool

The Science of Science (Sci2) Tool is a modular toolset specifically designed for the study of Science (Sci2 Team, 2009). Sci2 can be downloaded from http://sci2.cns.iu.edu and it can be freely used for research and teaching but also for commercial purposes (Apache 2.0 license). Extensive documentation is provided on the Sci2 Wiki (http://sci2.wiki.cns.iu.edu), or in the Information Visualization MOOC (http://ivmooc.cns.iu.edu), and the *Visual Insights* textbook (Börner & Polley, 2014).

Instead of focusing on just one specific type of analysis, like many other tools, it supports the temporal, geospatial, topical, and network analysis and visualization of datasets at the micro (individual), meso (local), and macro (global) levels. The tool is built on the OSGi/CIShell framework that is widely used and supported by industry (http://osgi.org). It uses an approach known as "plug-and-play" in macroscope construction (Börner, 2011), allowing anyone to easily add new algorithms and tools using a wizard-supported process, and to customize the tool to suit their specific research needs. The tool is optimized for datasets of up to 100,000 records for most algorithms (Light, Polley, & Börner, 2014). This section reviews key functionality—from data reading to preprocessing, analysis, and visualization.

14.2.1 Workflow Design and Replication

The tool supports workflow design, i.e., the selection and parameterization of different data readers, data analysis, visualization and other algorithms, via a unified graphical interface, see Fig. 14.1.



```
Russell J. Duhon. It uses the Cyberinfrastructure Shell (http://cishell.org) developed at the Cyberinfrastructure for Network Science Center
(http://cns.iu.edu) at Indiana University. Many algorithm plugins were derived from the Network Workbench Tool (http://nwb.cns.iu.edu)

Please cite as follows:
Sci2 Team. (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strategies, http://sci2.cns.iu.edu.
..........
Load... was selected.
Documentation: http://wiki.cns.iu.edu/display/CISHELL/Data+Formats
..........
Load... was selected.
Documentation: http://wiki.cns.iu.edu/display/CISHELL/Data+Formats
Found old-style ISI/Web Of Knowledge file.
Found old-style ISI/Web Of Knowledge file.
Found old-style ISI/Web Of Knowledge file.
Found old-style ISI/Web Of Knowledge file.
The original 361 records have been processed to remove duplicate unique ISI IDs leaving 361 records.

Wrote log to D:\Users\dapolley\AppData\Local\Temp\isiduplicateremoverlog2121540657214994626.txt
Loaded: D:\Users\dapolley\Desktop\sci2 v1.1\sampledata\scientometrics\isi\FourNetSciResearchers.isi
..........
Slice Table by Time was selected.
Implementer(s): Russell Duhon
```

- ISI Data: D:\Users\dapolley\Desktop\sci2 v1.1\sampl
  - 361 Unique ISI Records
    - Extracted Co-Authorship Network
    - Author information
- NSF csv file: D:\Users\dapolley\Desktop\sci2 v1.1\sa
  - visualized with Temporal Bar Graph
  - bar sizes

!	Algorithm Name	Date	Time	% Complete
☑	Temporal Bar Graph	11/26/2013	10:07:00 AM	🟩
☑	Load...	11/26/2013	10:06:45 AM	🟩
☑	Extract Co-Author Netw...	11/26/2013	10:06:39 AM	🟩
☑	Slice Table by Time	11/26/2013	09:28:34 AM	🟩
☑	Load...	11/26/2013	09:28:22 AM	🟩
☑	Load...	11/26/2013	09:28:16 AM	🟩

Fig. 14.1 Sci2 user interface with Menu on top, Console below, Scheduler in lower left, and Data Manager and Workflow Tracker on right.

Workflows are recorded and can be re-run to replicate results. They can be run with different parameter values to test and understand the sensitivity of results or to perform parameter sweeps. Workflows can also be run on other datasets to compare results. Last but not least, the same workflow can be run with different algorithms (e.g., clustering techniques) in support of algorithm comparisons. For a detailed documentation of workflow log formats and usage, please see our documentation wiki.[2]

[2] http://wiki.cns.iu.edu/display/CISHELL/Workflow+Tracker

All workflows discussed in this paper have been recorded and can be re-run. They can be downloaded from section 2.6, Sample Workflows, on the Sci2 wiki.[3]

14.2.2 Data Readers

The Sci2 Tool reads a number of common file formats, including tabular formats (.csv); output file formats from major data providers such as Thomson Reuter's Web of Science (.isi), Elsevier's Scopus (.scopus), Google Scholar, but also funding data from the U.S. National Science Foundation (.nsf) and the U.S. National Institutes of Health (using .csv); output formats from personal bibliography management systems such as EndNote (.endnote) and Bibtex (.bibtex). In addition, there exist data readers that retrieve data from Twitter, Flickr, and Facebook.[4] Last but not least, Sci2 was co-developed with the Scholarly Database (SDB) (http://sdb.cns.iu.edu) that provides easy access to 27 million paper, patent, grant, clinical trials records. All datasets downloaded from SDB in tabular or network format, e.g., co-author, co-inventor, co-investigator, patent-citation networks, are fully compatible with Sci2. File format descriptions and sample data files are provided at the Sci2 Wiki in Section 4.2.[5]

14.2.3 Temporal Analysis (When)

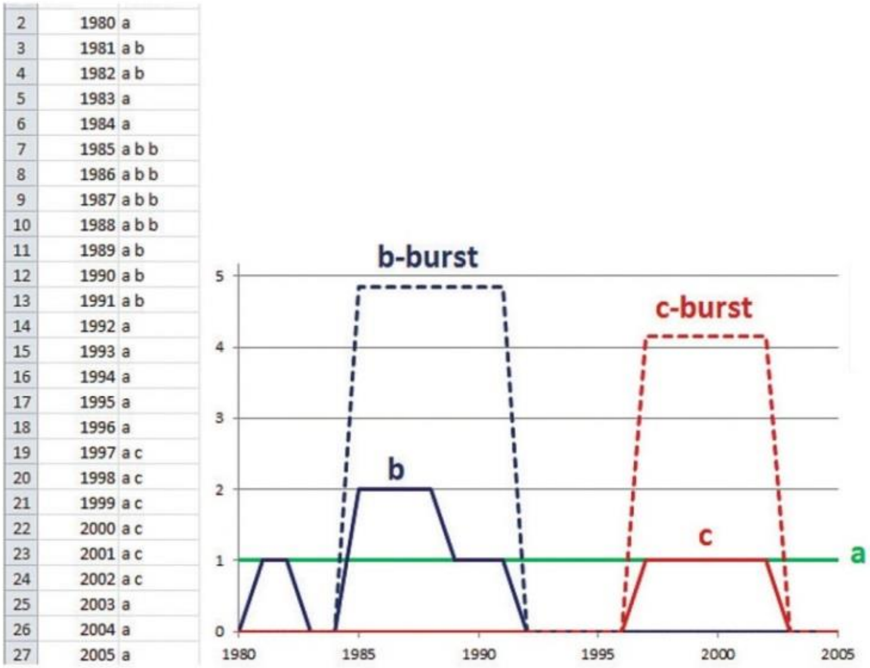
Data Preprocessing

The 'Slice Table by Time' algorithm[6] is a common data preprocessing step in many temporal visualization workflows. As an input, the algorithm takes a table with a date/time value associated with each record. Based on a user-specified time interval the algorithm divides the original table into a series of new tables. Depending on the parameters selected, these time slices are either cumulative or not, and aligned with the calendar or not. The intervals into which a table may be sliced include: milliseconds, seconds, minutes, hours, days, weeks, months, quarters, years, decades, and centuries.

Data Analysis

The 'Burst Detection' algorithm[7] implemented in Sci2, adapted from Jon Kleinberg's (2002), identifies sudden increases or "bursts" in the frequency-of-use of character strings over time. It identifies topics, terms, or concepts important to the events being studied that increased in usage, are more active for a period of time, and then fade away. The input for the algorithm is time-stamped text, such as documents with publication years. From titles, abstracts or other text, the algorithm generates a list of burst words, ranked according to burst weight, and the intervals of time in which these bursts occurred. Fig. 14.2 shows a diagram of bursting letters (left) next to the raw data (right). The letter 'b' (solid blue line plots frequency, dashed blue line plots burst) experienced a burst from just before 1985 to just after 1990. Similarly, the letter 'c' (red lines) experienced a burst starting just after 1995 and ending just before 2005. However, the letter 'a' (green line) remains constant throughout this time series, i.e., there is no burst for that term.

[3] http://wiki.cns.iu.edu/display/SCI2TUTORIAL/2.6+Sample+Workflows
[4] http://wiki.cns.iu.edu/display/SCI2TUTORIAL/3.1+Sci2+Algorithms+and+Tools
[5] http://wiki.cns.iu.edu/display/SCI2TUTORIAL/4.2+Data+Acquisition+and+Preparation
[6] http://wiki.cns.iu.edu/display/CISHELL/Slice+Table+by+Time
[7] http://wiki.cns.iu.edu/display/CISHELL/Burst+Detection



2	1980	a
3	1981	a b
4	1982	a b
5	1983	a
6	1984	a
7	1985	a b b
8	1986	a b b
9	1987	a b b
10	1988	a b b
11	1989	a b
12	1990	a b
13	1991	a b
14	1992	a
15	1993	a
16	1994	a
17	1995	a
18	1996	a
19	1997	a c
20	1998	a c
21	1999	a c
22	2000	a c
23	2001	a c
24	2002	a c
25	2003	a
26	2004	a
27	2005	a

Fig. 14.2 A burst analysis diagram for three letters (right) compared with the raw data (left).

Data Visualization

The ‘Temporal Bar Graph’ visualizes numeric data over time, and is the only truly temporal visualization algorithm available in Sci2.[8] This algorithm accepts tabular (CSV) data, which must have start and end dates associated with each record. Records that are missing either start or end dates are ignored. The other input parameters include ‘Label’, which corresponds to a text field and is used to label the bars; ‘Size By’, which must be an integer and corresponds to the area of the horizontal bars; ‘Date Format’, which can either be in “Day-Month-Year Date Format (Europe, e.g. 31/10/2010)” or “Month-Day-Year (U.S., e.g. 10/31/2010)”; and ‘Category’, which allows users to color code bars by an attribute of the data. For example, Fig. 14.3 shows the National Science Foundation (NSF) funding Profile for Dr. Geoffrey Fox, Associate Dean for Research and Distinguished Professor of Computer Science and Informatics at Indiana University, where each bar represents an NSF award on which Dr. Fox was an investigator. Each bar is labeled with the title of the award. The total area of the bars corresponds to the total amount awarded, and the bars are color-coded by the institution/organization affiliated with the award.

[8] http://wiki.cns.iu.edu/display/CISHELL/Temporal+Bar+Graph

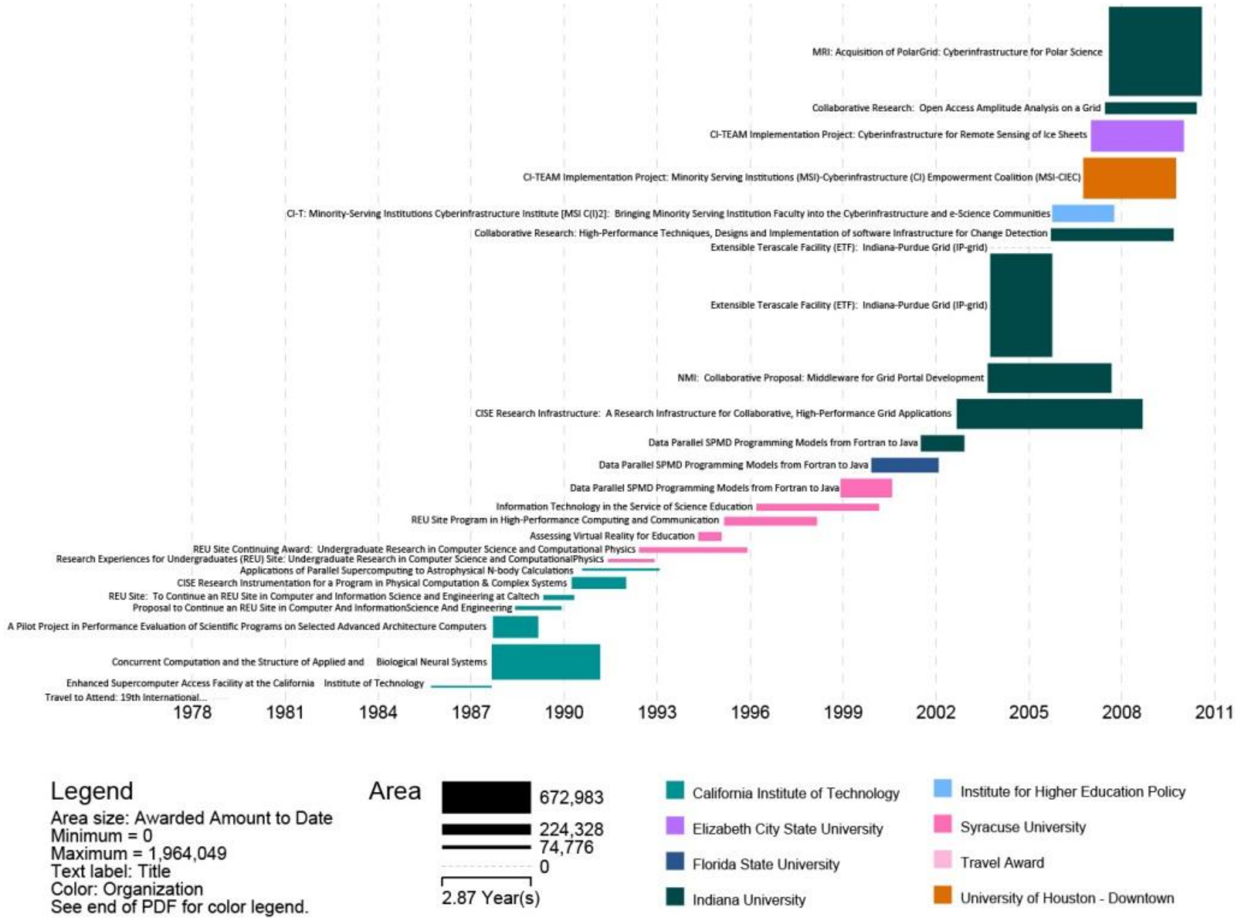


Fig. 14.3 Temporal Bar Graph showing the NSF funding profile for Dr. Geoffrey Fox

14.2.4 Geospatial Analysis (Where)

Data Preprocessing

The 'Extract ZIP Code' algorithm[9] is a common preprocessing step in geospatial visualization workflows. This algorithm takes US addresses as input data and extracts the ZIP code from the address in either the standard 5-digit short form (xxxxx) or the standard 9-digit long form (xxxxx-xxxx). This feature facilitates quick spatial analysis and simplifies geocoding. However, the algorithm is limited to U.S. or U.S.-based ZIP code systems. Another useful preprocessing step for geospatial and other workflows is data aggregation. Redundant geo-identifiers are common in geospatial analysis and require aggregation prior to visualization. Sci2 provides basic aggregation with the 'Aggregate Data' algorithm[10], which groups together values in a column selected by the user. The other values in the records are aggregated as specified by the user. Currently, sum, difference, average, min, and max are available for numerical data. All text data are aggregated when a text delimiter is provided.

Data Analysis

[9] http://wiki.cns.iu.edu/display/CISHELL/Extract+ZIP+Code
[10] http://wiki.cns.iu.edu/display/CISHELL/Aggregate+Data

Sci2 has a variety of geocoding options. The ‘Generic Geocoder’[11] is the most basic of these options. It converts U.S. addresses, U.S. states, and U.S. ZIP codes into longitude and latitude values. The input for this algorithm is a table with a geo-identifier for each record, and the output is the same table but with a longitude and latitude value appended to each record. There are no restrictions on the number of records that can be geocoded using the ‘Generic Geocoder’. The ‘Bing Geocoder’[12] expands the functionality of the ‘Generic Geocoder’, allowing Sci2 to convert international addresses into longitude and latitude values. All coordinates are obtained by querying the Bing geocoder service and Internet access must be available while using this algorithm. Users must obtain an API key from Bing Maps in order to run this algorithm, and there is a limit of 50,000 records which can be geocoded in a 24 hour period. Finally, Sci2 provides a ‘Congressional District Geocoder’[13], which converts 9-digit ZIP codes (5-digit ZIP codes can contain multiple districts) into congressional districts and geographic coordinates. The algorithm is available as an external plugin that can be downloaded from the Sci2 wiki.[14] The database that supports this algorithm is based on the 2012 ZIP Code Database for the 113th U.S. Congress and does not take into account any subsequent redistricting.

Data Visualization

The Sci2 Tool offers three geospatial visualization algorithms: a proportional symbol map, a choropleth map, or region shaded map, and a network geomap overlay. The ‘Proportional Symbol Map’ algorithm[15] takes a list of coordinates and at most three numeric attributes and visualizes them over a world or United States base map. The sizes and colors of the symbols are proportional to the numeric values. The ‘Choropleth Map’ algorithm[16] allows users to color countries of the world or states of the United States in proportion to one numeric attribute. Finally, Sci2 has the ability to visualize networks overlaid on a geospatial basemap. As input, the ‘Geospatial Network Layout with Base Map’ algorithm[17] requires a network file with latitude and longitude values associated with each node. The algorithm produces a network file and a PostScript base map. The network file is visualized in GUESS or Gephi, exported as a PDF, and overlaid on the PDF of the base map. Fig. 14.4 shows an exemplary proportional symbol map (left), choropleth map (center), and the geospatial network layout with base map (right).

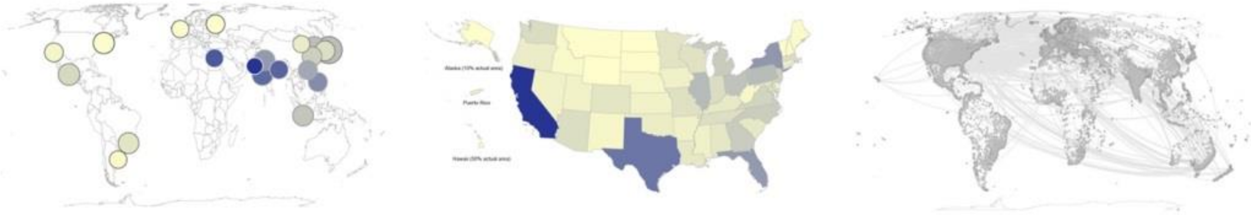


Fig. 14.4 Proportional symbol map (left), choropleth map (center), and network layout overlaid on world map (right)

14.2.5 Topical Analysis (What)

Data Preprocessing

[11] http://wiki.cns.iu.edu/display/CISHELL/Geocoder
[12] http://wiki.cns.iu.edu/display/CISHELL/Bing+Geocoder
[13] http://wiki.cns.iu.edu/display/CISHELL/Congressional+District+Geocoder
[14] http://wiki.cns.iu.edu/display/SCI2TUTORIAL/3.2+Additional+Plugins
[15] http://wiki.cns.iu.edu/display/CISHELL/Proportional+Symbol+Map
[16] http://wiki.cns.iu.edu/display/CISHELL/Choropleth+Map
[17] http://wiki.cns.iu.edu/display/CISHELL/Geospatial+Network+Layout+with+Base+Map

The topic or semantic coverage of a scholar, institution, country, paper, journal or area of research can be derived from the texts associated with it. In order to analyze and visualize topics, text must first be normalized. Sci2 provides basic text normalization with the 'Lowercase, Tokenize, Stem, and Stopword Text' algorithm.[18] This algorithm requires tabular data with a text field as input and outputs a table with the specified text field normalized. Specifically, the algorithm makes all text lowercase, splits the individual words into tokens (delimited by a user-selected separator), stems each token (removing low content prefixes and suffixes), and removes stopwords, i.e., very common (and therefore dispensable) words or phrases such as "the" or "a". Sci2 provides a basic stopword list[19], which can be edited to fit users' specific needs. The goal of text normalization is to facilitate the extraction of unique words or word profiles in order to identify topic coverage of bodies of text.

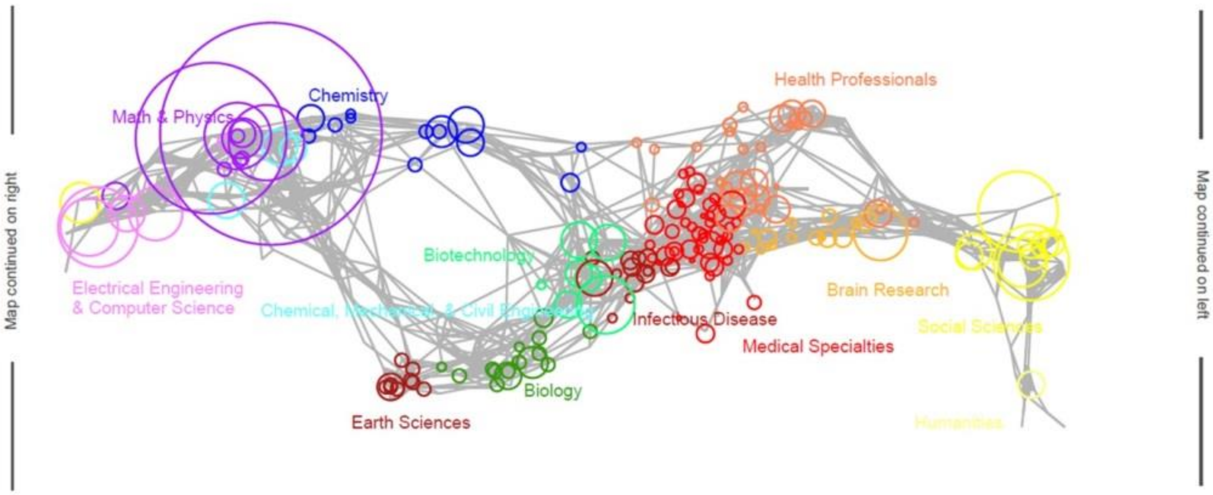
Data Analysis

Burst detection, previously discussed in the temporal section, is often used for identifying the topic coverage of a corpus of text. Since burst detection also involves a temporal component, it is ideal for demonstrating the evolution of scientific research topics as represented by bodies of text.

Data Visualization

Sci2 provides a map of science visualization algorithm to display the topical distributions, also called expertise profiles. The UCSD Map of Science (Börner et al., 2012) is a visual representation of 554 sub-disciplines within the 13 disciplines of science and their relationships to one another. There are two variations of this algorithm: 'Map of Science via Journals' and 'Map of Science via 554 Fields'.[20,21] The first works by matching journal titles to the underlying sub-disciplines, as specified in the UCSD Map of Science classification scheme.[22] The second works by directly matching the IDs for the 554 fields, integers 1 to 554, to the sub-disciplines. Both algorithms take tabular data as input, the first with a column of journal names and the second with a column of field IDs. It is recommended that users run the 'Reconcile Journal Names' algorithm[23] prior to science mapping. Both map of science algorithms output a PostScript file and the 'Map of Science via Journals' also outputs two tables: one for the journals located, and one for journals not located. Fig. 14.5 shows the topic distribution of the *FourNetSciResearchers.isi* file, a dataset containing citations from four major network science researchers: Eugene Garfield, Stanley Wasserman, Alessandro Vespignani, and Albert-László Barabási with a total of 361 publication records.

[18] http://wiki.cns.iu.edu/display/CISHELL/Lowercase%2C+Tokenize%2C+Stem%2C+and+Stopword+Text
[19] ***yoursci2directory****/configuration/stopwords.txt*
[20] http://wiki.cns.iu.edu/display/CISHELL/Map+of+Science+via+Journals
[21] http://wiki.cns.iu.edu/display/CISHELL/Map+of+Science+via+554+Fields
[22] http://sci.cns.iu.edu/ucsdmap
[23] http://wiki.cns.iu.edu/display/CISHELL/Reconcile+Journal+Names



2008 The Regents of the University of California and SciTech Strategies.
Map updated by SciTech Strategies, OST, and CNS in 2011.

Fig. 14.5 Map of Science via Journals showing the topic coverage of the *FourNetSciResearchers.isi* file, see wiki for legend and additional information analogous to Fig. 14.8.

14.2.6 Network Analysis (With Whom)

Data Preprocessing

Many datasets come in tabular format and a key step in most network visualization workflows involves extracting networks from these tables. Sci2 supports this process with a variety of network extraction algorithms. The 'Extract Co-Occurrence Network' algorithm can be used to extract networks from columns that contain multiple values.[24] The 'Extract a Directed Network' algorithm[25] will create a network between two columns with data of the same type. Sci2 makes extracting co-occurrence networks specific to bibliometric analysis even easier by providing algorithms such as 'Extract Co-Author Network', 'Extract Word Co-Occurrence Network', and 'Extract Reference Co-Occurrence (Bibliographic Coupling) Network'.[26, 27, 28] Finally, for two columns that contain different data types, the 'Extract Bipartite Network' algorithm[29] can be used.

Data Analysis

Sci2 offers a large variety of network analysis algorithms for directed or undirected, and weighted or unweighted networks. A full list of all available network analysis algorithms can be found in section 3.1 Sci2 Algorithms and Tools of the online wiki.[30] Two of the more interesting algorithms for network analysis include the PageRank and Blondel Community Detection algorithms (also known as the Louvain algorithm). The PageRank algorithm was originally developed for the Google search engine to rank sites in the search

[24] http://wiki.cns.iu.edu/display/CISHELL/Extract+Word+Co-Occurrence+Network
[25] http://wiki.cns.iu.edu/display/CISHELL/Extract+Directed+Network
[26] http://wiki.cns.iu.edu/display/CISHELL/Extract+Co-Author+Network
[27] http://wiki.cns.iu.edu/display/CISHELL/Extract+Word+Co-Occurrence+Network
[28] http://wiki.cns.iu.edu/display/CISHELL/Extract+Reference+Co-Occurrence+%28Bibliographic+Coupling%29+Network
[29] http://wiki.cns.iu.edu/display/CISHELL/Extract+Bipartite+Network
[30] http://wiki.cns.iu.edu/display/SCI2TUTORIAL/3.1+Sci2+Algorithms+and+Tools

result by relative importance, as measured by the number of links to a page (Brin & Page, 1998).[31] The same process can be used in directed networks to rank the relative importance of nodes. There are two versions of the 'PageRank' algorithm in Sci2, one for directed and unweighted networks, which simply measures the importance of nodes based on the number of incoming edges, and one for directed and weighted networks, which measures the importance nodes based on incoming edges and takes into consideration the weight of those edges. Both algorithms are useful for identifying important nodes in very large networks. The 'Blondel Community Detection' algorithm is a clustering algorithm for large networks.[32] The algorithm detects communities in weighted networks using an approach based on modularity optimization (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008). The resulting network will be structurally the same but each node will have an attribute labeled "blondel_community_level_x."

Data Visualization

The Sci2 Tool offers multiple ways to view networks. While the tool itself does not directly support network visualization, GUESS, a common network visualization program comes already bundled with the tool (Adar & Kim, 2007). For users who already have Gephi (Bastian et al., 2009) installed on their machines, Sci2 provides a bridge, allowing a user to select a network in the Data Manager, then run Gephi, and the tool will start with the selected network file loaded. Gephi is available for download at http://gephi.org. Cytoscape (Saito et al., 2012) was also made available as a plugin to the Sci2 tool, further expanding network visualization options. The Cytoscape plugin is available for download from section 3.2 of the Sci2 wiki.[33]

14.3 Career Trajectories

This section demonstrates Sci2 Tool functionality for the analysis and visualization of career trajectories, i.e., the trajectory of people, institutions, or countries over time. In addition to studying movement over geospatial space, e.g., via the addresses of different institutions people might study and work at, the expertise profile of one physicist is analyzed. Specifically, we use a dataset of authors in physics with a large number of institutional affiliations—assuming that those authors engaged in a large number of Postdocs. The original dataset of the top-10,000 authors in physics with the most affiliations was retrieved by Vincent Larivière. For the purposes of this study, we selected 10 of top paper producing authors. Using the name-unification method introduced by Kevin W. Boyack and Richard Klavans (2008), we consider names to refer to one person if the majority of his or her publications occur at one institution. Uniquely identifying people in this way is imperfect but practical. The resulting dataset contains 10 authors that have more than 70 affiliation addresses in total and published in more than 100 mostly physics journals between 1988 and 2010. Subsequently, we show how Sci2 can be used to clean and aggregate data, provide simple statistics, analyze the geospatial and topical evolution of the different career trajectories over time and visualize results. We conclude this section with an interpretation of results and a discussion of related works.

14.3.1 Data Preparation Analysis

The list of the top 10,000 physicists was loaded into Google Refine for cleaning.[34] The author names were normalized by placing them in uppercase and trimming leading and trailing white spaces. Then, the text facets feature of Google Refine was used to identify groups of names. The count associated with each name corresponds to the number of papers associated with that name, which starts to give some idea which names

[31] http://wiki.cns.iu.edu/display/CISHELL/PageRank
[32] http://wiki.cns.iu.edu/display/CISHELL/Blondel+Community+Detection
[33] http://wiki.cns.iu.edu/display/SCI2TUTORIAL/3.2+Additional+Plugins
[34] http://code.google.com/p/google-refine/

uniquely identify a person, reducing homonymy. Once a name was identified as potentially uniquely identifying a person, the text facet was applied to this person's institutions, showing the number of papers produced at each institution. If the majority of the papers associated with a name occurred at one institution, the name was considered to uniquely identify one person in this dataset. Following this process for the highest producing authors in this dataset resulted in a list of 10 physicists, see Table 14.1.

Table 14.1 Top-10 physicists with the most publications plus the number of their institutions, papers, and citation counts

Name	Institutions	Papers	Citations
AGARWAL-GS	7	163	3917
AMBJORN-J	3	185	3750
BENDER-CM	6	118	3962
BRODSKY-SJ	14	119	4566
CHAICHIAN-M	7	123	2725
ELIZALDE-E	16	135	2151
GIANTURCO-FA	4	130	1634
PERSSON-BNJ	6	100	3472
YUKALOV-VI	8	150	1772
ZHDANOV-VP	6	147	1594

Next, the addresses for each institution were obtained by searching the Web. The data for each of the 10 authors was then saved in separate CSV files and loaded into Sci2. The 'Bing Geocoder' algorithm[35] was used to geocode each institution. Then, the 'Aggregate Data' algorithm[36] was used to aggregate the data for each author by institution, summing the number of papers produced at those institutions and summing the total citations for those papers. This aggregation was performed because many authors had affiliations with institutions far away from their home institutions for short periods of time, either due to sabbaticals or as visiting scholars.

In addition, the citation data for each of the 10 physicists was downloaded from the Web of Science. The full records plus citations were exported as ISI formatted text files. The extensions were changed from .txt to .isi and loaded into Sci2. Next, the 'Reconcile Journal Names' algorithm[37] was run, which ensures all journal titles are normalized and matched to the standard given in the UCSD Map of Science Standard (Börner et al., 2012).

14.3.2 Data Visualization and Interpretation

Initially a comparison of all the authors, their paper production, and resulting citation counts was created in MS Excel. The 10 authors followed more or less similar trajectories, with earlier works receiving more citations and newer works fewer. Fig. 14.6 shows the number of papers per publication year (blue line) and the number of all citations received up to 2009 by the papers published in a specific year (red line). As expected, the bulk of the paper production tends to happen at one "home" institution.

[35] http://wiki.cns.iu.edu/display/CISHELL/Bing+Geocoder
[36] http://wiki.cns.iu.edu/display/CISHELL/Aggregate+Data
[37] http://wiki.cns.iu.edu/display/CISHELL/Reconcile+Journal+Names

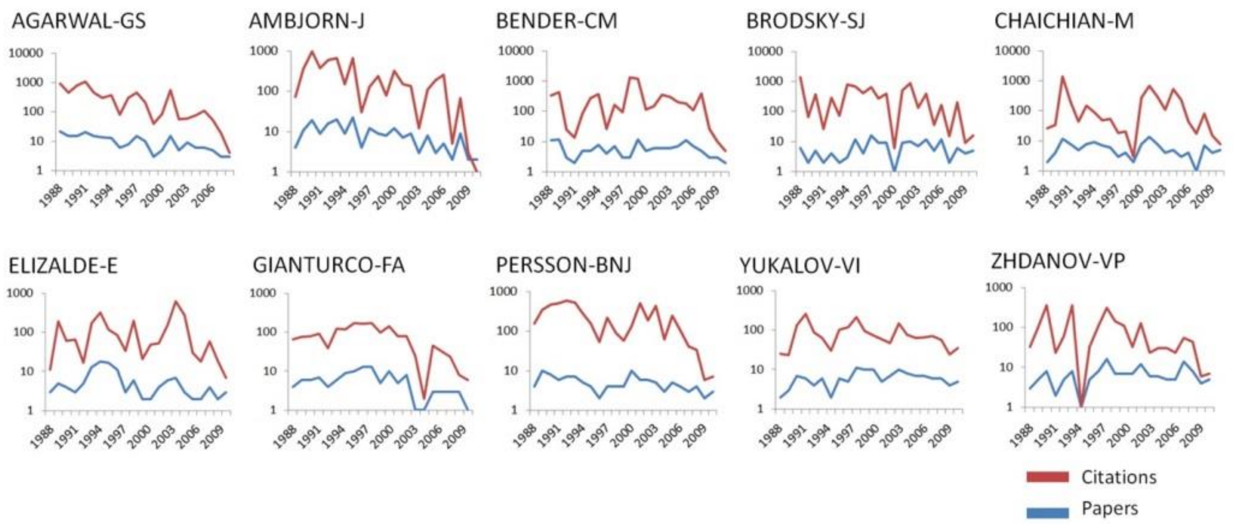


Fig. 14.6 Paper and citation counts over 23 years for each of the 10 physicists

Next, the geocoded files for each author were visualized in Sci2 to show the world-wide career trajectories over geospatial space. Specifically, the 'Proportional Symbol Map' algorithm[38] was applied to the physicist with the highest combined number of papers and citation count and the resulting map was saved as PDFs and further edited in Adobe Illustrator to add the directed edges connecting the institution symbols. The resulting map for Dr. Girish Agarwal is shown in Fig. 14.7. Each circle corresponds to an institution and is labeled with the institution's name and the date range in years that the physicist is associated with the institution. The symbols are sized proportional to the number of papers produced at each institution and colored proportional to the number of citations those papers have received. A green "Start" label was added to the first institution associated with the physicist in the dataset, and a red "End" label was added to the last. Green numbers indicate the sequence of transitions between start, intermediary, and end institutions.

[38] http://wiki.cns.iu.edu/display/CISHELL/Proportional+Symbol+Map

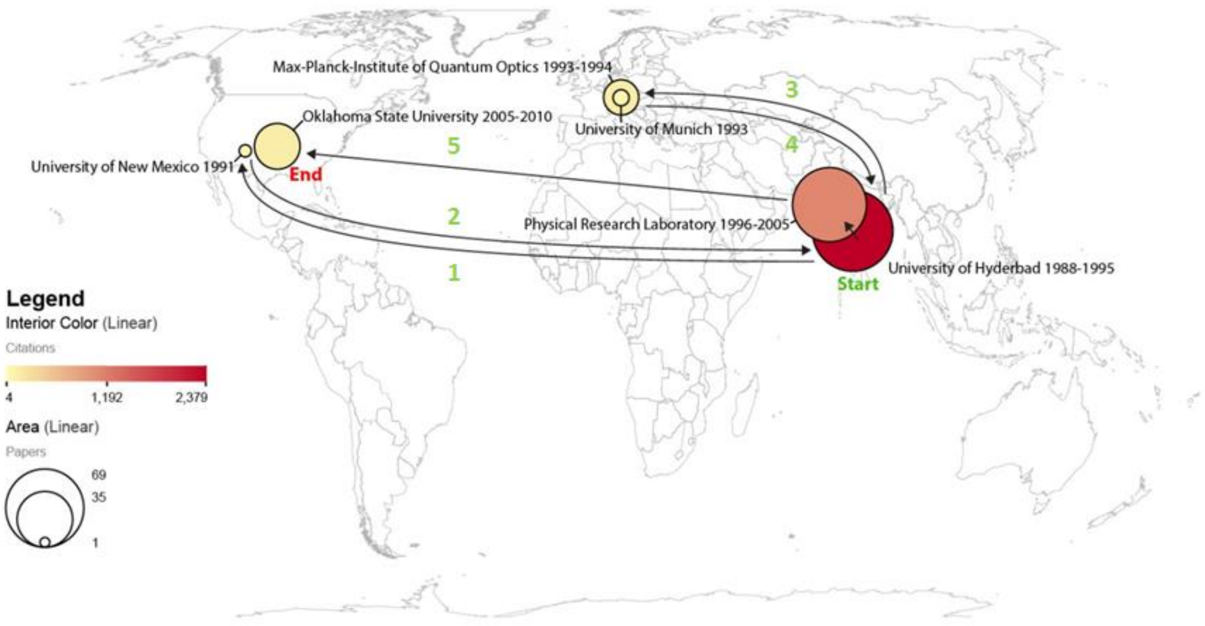
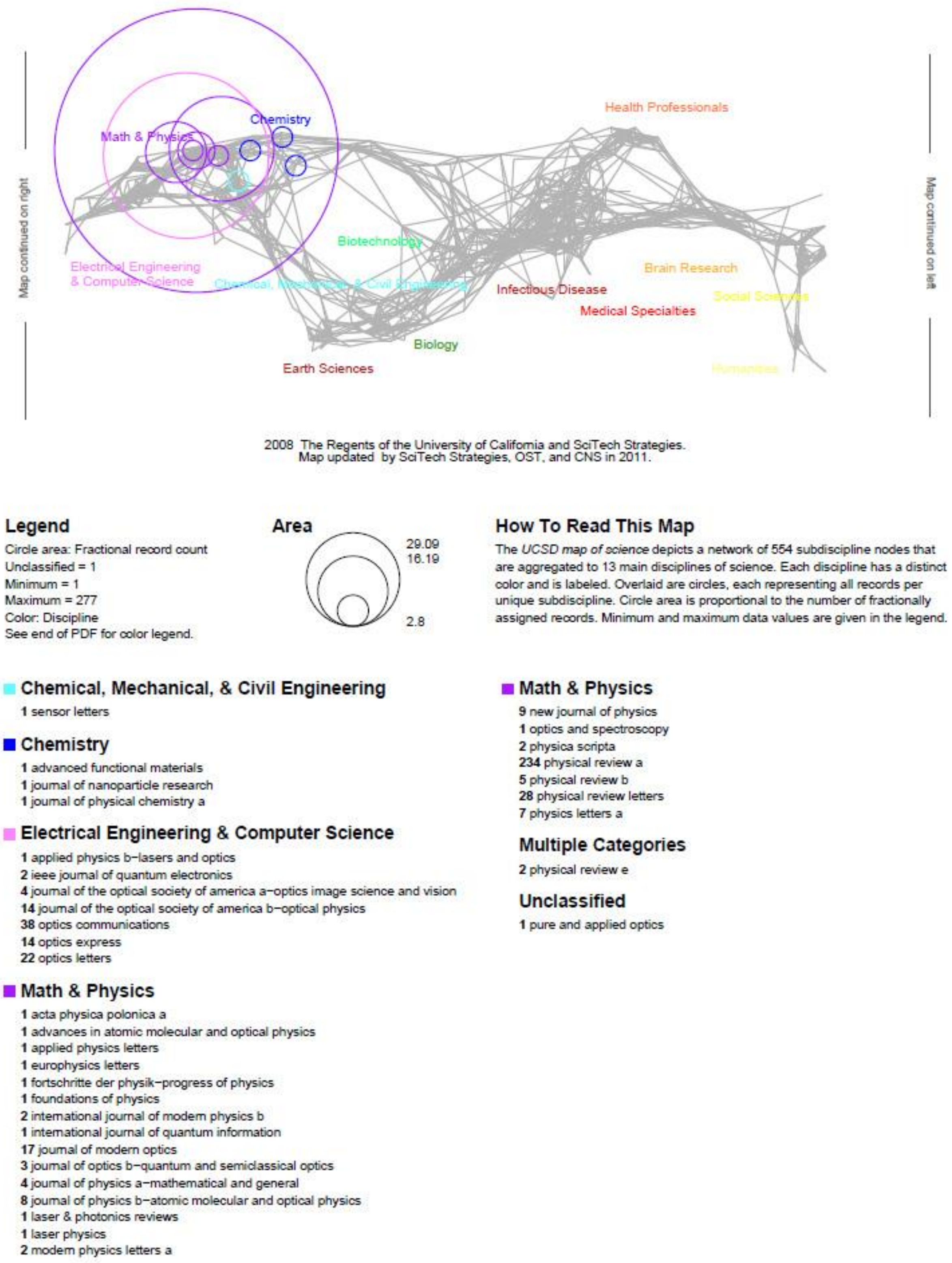

Max-Planck-Institute of Quantum Optics 1993-1994
University of Munich 1993
Oklahoma State University 2005-2010
University of New Mexico 1991
End
Physical Research Laboratory 1996-2005
University of Hyderbad 1988-1995
Start
1
2
3
4
5
Legend
Interior Color (Linear)
Citations
4
1,192
2,379
Area (Linear)
Papers
69
35
1

Fig. 14.7 Career trajectory for Dr. Girish Agarwal, 1988-2010

As expected, the majority of the paper production occurred at the institutions where Dr. Agarwal spent the most time, University of Hyderbad and Physical Research Laboratory, where he was the director for 10 years. The papers produced at these institutions also have the highest citations counts, as these papers are much older than the others in the dataset. Visualizing the dataset over a geographic map gives a sense of the geospatial trajectory of Agarwal's career, but the limited scope of the dataset results in a somewhat misleading visualization. Simply by looking at the map, one might assume that Dr. Agarwal started his career at Hyderbad University, but he actually received his Ph.D. from the University of Rochester in 1969. It is highly likely there are other publications at other institutions that, due to the limited date range, are not captured by this visualization.

Next, to visualize the topical distribution Dr. Agarwal's publications, his ISI file was mapped using the 'Map of Science via Journals' using the default parameter values. Fig. 14.8 shows the topic distribution of Dr. Agarwal. As expected, the majority of his publications occur in the fields of Math & Physics and Electrical Engineering & Computer Science.



2008 The Regents of the University of California and SciTech Strategies.
Map updated by SciTech Strategies, OST, and CNS in 2011.

Legend

Circle area: Fractional record count
Unclassified = 1
Minimum = 1
Maximum = 277
Color: Discipline
See end of PDF for color legend.

Area

How To Read This Map

The *UCSD map of science* depicts a network of 554 subdiscipline nodes that are aggregated to 13 main disciplines of science. Each discipline has a distinct color and is labeled. Overlaid are circles, each representing all records per unique subdiscipline. Circle area is proportional to the number of fractionally assigned records. Minimum and maximum data values are given in the legend.

■ Chemical, Mechanical, & Civil Engineering

1 sensor letters

■ Chemistry

1 advanced functional materials
1 journal of nanoparticle research
1 journal of physical chemistry a

■ Electrical Engineering & Computer Science

1 applied physics b−lasers and optics
2 ieee journal of quantum electronics
4 journal of the optical society of america a−optics image science and vision
14 journal of the optical society of america b−optical physics
38 optics communications
14 optics express
22 optics letters

■ Math & Physics

1 acta physica polonica a
1 advances in atomic molecular and optical physics
1 applied physics letters
1 europhysics letters
1 fortschritte der physik−progress of physics
1 foundations of physics
2 international journal of modern physics b
1 international journal of quantum information
17 journal of modern optics
3 journal of optics b−quantum and semiclassical optics
4 journal of physics a−mathematical and general
8 journal of physics b−atomic molecular and optical physics
1 laser & photonics reviews
1 laser physics
2 modern physics letters a

■ Math & Physics

9 new journal of physics
1 optics and spectroscopy
2 physica scripta
234 physical review a
5 physical review b
28 physical review letters
7 physics letters a

Multiple Categories

2 physical review e

Unclassified

1 pure and applied optics

Fig. 14.8 Topic distribution of Dr. Agarwal's ISI publications on the Map of Science and discipline specific listing of journals in which he published. Note the five journals in 'Unclassified' that could not be mapped.

Finally, to analyze the connections between the ten different physicists and the 112 journals in which they publish, a bipartite network was created. A property file was used to add the total citation counts to each node. Fig. 14.9 shows the graph of all ten physicists, where the circular nodes represent the authors and the square nodes represent the journals. The author nodes are sized according to their out-degree, or the number of journals in which they publish, and the journal nodes are sized by their in-degree, or more popular publication venues within the context of this dataset. The nodes are colored from yellow to red based on citation count, with all author nodes labeled and the journal nodes with the highest citation counts also labeled.

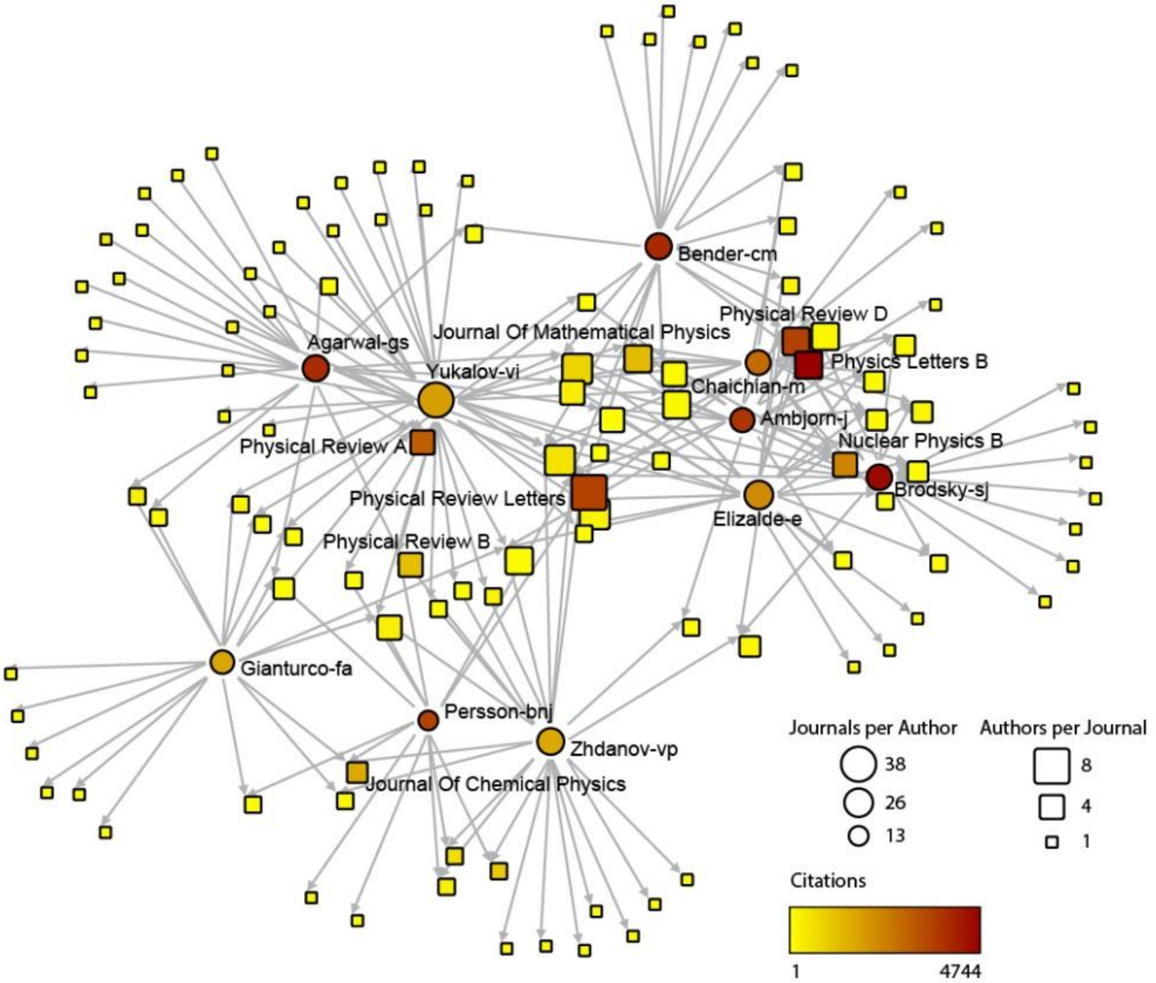


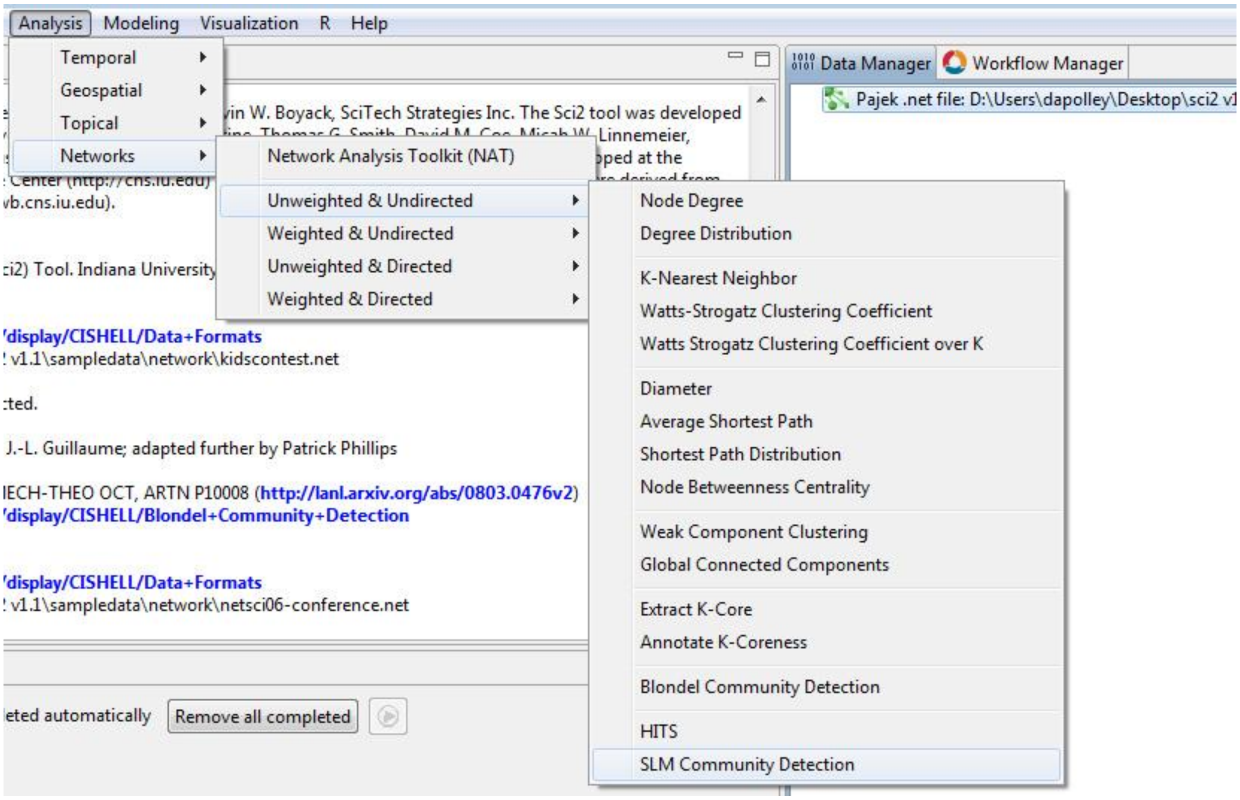
Fig. 14.9 Bi-modal network of all ten authors and the 112 journals in which they publish

The author with the most diverse publication venues is Vyacheslav Yukalov who published in 38 unique journals. The author with the highest citation count in this dataset is Girish Agarwal, with 3,917 citations to 163 papers. The journal with publications by the greatest number of authors in this dataset is *Physical Review B*, but the journal with papers that have the highest total citation count is *Physics Letters B*, with 4,744 citations.

14.4 Discussion and Outlook

The Sci2 Tool is one of several tools that use the OSGi/CIShell framework (Börner, 2011, 2014). Other tools comprise the Network Workbench (NWB) designed for advanced network analysis and visualization (http://nwb.cns.iu.edu); the Epidemiology tool (EpiC) that supports model building and real time analysis of data and adds a bridge to the R statistical package (http://epic.cns.iu.edu); and TexTrend for textual analysis (http://textrend.org) (Kampis et al., 2009). Thanks to the unique plug-and-play macroscope approach, plugins can be shared between the different tools, allowing individuals to customize the tool to suit their specific research needs.

Current work focuses on the integration of the smart local moving (SLM) algorithm (Waltman & Eck, 2013) into Sci2 making it possible to run clustering on any network file (Fig. 14.10). This algorithm detects communities based on the relative number of links to nodes in a network. Future development will incorporate the full community detection capabilities of VOSviewer, which allows users to specify the level of granularity for clustering, resulting in fewer larger communities or more smaller communities (Van Eck & Waltman, 2010).



vin W. Boyack, SciTech Strategies Inc. The Sci2 tool was developed

Linnemeier,

oped at the

: Center (http://cns.iu.edu)

vb.cns.iu.edu).

ci2) Tool. Indiana University

'display/CISHELL/Data+Formats

! v1.1\sampledata\network\kidscontest.net

cted.

J.-L. Guillaume; adapted further by Patrick Phillips

IECH-THEO OCT, ARTN P10008 (http://lanl.arxiv.org/abs/0803.0476v2)

'display/CISHELL/Blondel+Community+Detection

'display/CISHELL/Data+Formats

! v1.1\sampledata\network\netsci06-conference.net

Fig. 14.10 Sci2 menu showing the SLM Community Detection algorithm

In addition, the Sci2 Tool can be run as a web service making it possible to request analyses and visualizations online and to execute more computing intensive jobs in the cloud. A first interface will soon be deployed at the National Institutes of Health RePORTER site. Eventually, Sci2 as a web service will be released publicly with full build support so that users can build and deploy Sci2 to the Web themselves. Furthermore, algorithms developed for the desktop version will be compatible with the online version. Ideally, users will be able to use the workflow tracker combined with the Sci2 web service functionality to create web applications that read in data and output visualizations.

14.5 Acknowledgments

We would like to thank Vincent Larivière for compiling the dataset of the top-10,000 physicists with the most affiliations and Robert P. Light and Michael P. Ginda for comments on an earlier draft of the paper. The Sci2 tool and web services were/are developed by Chin Hua Kong, Adam Simpson, Steven Corenflos, Joseph Biberstine, Thomas G. Smith, David M. Coe, Micah W. Linnemeier, Patrick A. Phillips, Chintan Tank, and Russell J. Duhon; the primary investigators were Katy Börner, Indiana University and Kevin W. Boyack, SciTech Strategies Inc. Sci2 uses the Cyberinfrastructure Shell (http://cishell.org) developed at the Cyberinfrastructure for Network Science Center (http://cns.iu.edu) at Indiana University. Many algorithm plugins were derived from the Network Workbench Tool (http://nwb.cns.iu.edu). This work was partially funded by the National Science Foundation under award SBE-0738111 and National Institutes of Health under award U01 GM098959.

References

Adar, E., & Kim, M. (2007). *SoftGUESS: Visualization and Exlporation of Code Clones in Context*. Paper presented at the 29th International Conference on Software Engineering, Minneapolis, MN.

Andrienko, N., Andrienko, G., Voss, H., Bernardo, F., Hipolito, J., & Kretchmer, U. (2002). Testing the Usability of Interactive Maps in CommonGIS. *Cartography and Geographic Information Science, 29*(4), 325-342. doi: 10.1559/152304002782008369

Anselin, L., Syabri, I., & Kho, Y. (2006). GeoDa: An Introduction to Spatial Data Analysis. *Geographical Analysis, 38*(1), 5-22.

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the Third International ICWSM Conference*, 361-362.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast Unfolding of Communities in Large Networks. *Journal of Statistical Mechanics*, P10008. doi: 10.1088/1742-5468/2008/10/P10008

Börner, K. (2011). Plug-and-Play Macroscopes. *Communications of the ACM, 54*(3), 60-69. doi: 10.1145/1897852.1897871

Börner, K. (2014). Plug and Play Macroscopes: Network Workbench (NWB), Science of Science Tool, (Sci2), and Epidemiology Tool (Epic) *Encyclopedia of Social Network Analysis and Mining*: Springer Verlag.

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., . . . Boyack, K. W. (2012). Design and Update of a Classifcation System: The UCSD Map of Science. *PLOS ONE, 7*(7), e39464-39464. doi: 10.1371/journal.pone.0039464

Börner, K., & Polley, D. E. (2014). *Visual Insights: A Practical Guide to Making Sense of Data*. Boston, MA: MIT Press.

Boyack, K. W., & Klavans, R. (2008). Measuring Science-Technology Interaction Using Rare Inventor-Author Names. *Journal of Infometrics, 2*(3), 173-182. doi: 10.1016/j.joi.2008.03.001

Brin, S., & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks and ISDN Systems, 30*(1-7), 107-118.

Buono, P., Plaisant, C., Simeone, A., Aris, A., Shneiderman, B., Shmueli, G., & Jank, W. (2007). *Similarity-Based Forecasting with Simultaneous Previews: A River Plot Interface for Time Series Forecasting*. Paper presented at the 11th International Conference Information Visualization (IV '07), Zurich, Switzerland.

Cronin, B., & Sugimoto, C. R. (Eds.). (2014). *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*. Cambridge, MA: MIT Press.

Hardeman, S., Hoekman, J., & Frenken, K. (2014). Spatial scientometrics: Taking stock, looking ahead. In Y. Ding, R. Rousseau & D. Wolfram (Eds.), *Measuring Scholarly Impact: Methods and Practice*: Springer.

Hochheiser, H., & Shneiderman, B. (2004). Dynamic Query Tools for Time Series Data Sets, Timebox Widgets for Interactive Exploration. *Information Visualization, 3*(1), 1-18.

Kampis, G., Gulyas, L., Szaszi, Z., Szakolczi, Z., & Soos, S. (2009). *Dynamic Social Networks and the TexTrend/CIShell Framework*. Paper presented at the Conference on Applied Social Network Analysis (ASNA): 21. http://pb.dynanets.org/publications/DynaSocNet_TexTrend_v2.0.pdf

Kleinberg, J. M. (2002). Bursty and Hierarchical Structure in Streams. *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 91-101. doi: 10.1145/775047.775062

Light, R., Polley, D., & Börner, K. (2014). Open data and open code for big science of science studies. *Scientometrics*, 1-17. doi: 10.1007/s11192-014-1238-2

Milojevic, S. (2014). Network property and dynamics. In Y. Ding, R. Rousseau & D. Wolfram (Eds.), *Measuring Scholarly Impact: Methods and Practice*: Springer.

Nooy, W. D., Mrvar, A., & Batageli, V. (2011). *Exploratory Social Network Analysis with Pajek* (2nd ed.): Cambridge University Press.

Saito, R., Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L., Lotia, S., . . . Ideker, T. (2012). A Travel Guide to Cytoscape Plugins. *Nature Methods*, *9*(11), 1069-1076.

Scharnhorst, A., Börner, K., & van den Besselaar, P. (Eds.). (2012). *Models of Science Dynamics: Encounters Between Complexity Theory and Information Science*. New York, NY: Springer.

Sci2 Team. (2009). Science of Science (Sci2) Tool: Indiana University and SciTech Strategies. Retrieved from http://sci2.cns.iu.edu

Song, M., & Chambers, T. (2014). Text mining In Y. Ding, R. Rousseau & D. Wolfram (Eds.), *Measuring Scholarly Impact: Methods and Practice*: Springer.

Takatsuka, M., & Gahegan, M. (2002). GeoVISTA *Studio*: A Codeless Visual Programming Environment for Geoscientific Data Analysis and Visualization. *Computers & Geosciences, 28*(10), 1131-1144. doi: 10.1016/S 0098-3004(02)00031-6

Van Eck, N., & Waltman, L. (2010). Software Survey: VOSviewer, a Computer Program for Bibliometric Mapping. *Scientometrics*, *84*(2), 523-538.

Van Eck, N., & Waltman, L. (2014). Visualizing bibliometric data. In Y. Ding, R. Rousseau & D. Wolfram (Eds.), *Measuring Scholarly Impact: Methods and Practice*: Springer.

Waltman, L., & Eck, N. J. v. (2013). A Smart Local Moving Algorithm for Large-Scale Modularity-Based Community Detection. *European Physical Journal B*, *86*(11), 471-485.

Waltman, L., & Yan, E. (2014). PageRank-inspired bibliometric indices In Y. Ding, R. Rousseau & D. Wolfram (Eds.), *Measuring Scholarly Impact - Methods and Practice*: Springer.